WLCG/HSF Workshop, Live Notebook

This is the live notebook for the WLCG/HSF workshop in Naples. In the finest traditions of the do-ocracy it's open to everyone to take notes on the important points, add suggestions or ask questions. (Just please don't clobber the comments of others!)

Monday

Opening Plenary

HEP science goals for the next decade (Liz Sexton-Kennedy)

Summary of expected resource needs for LHC
Comparison with SKA ("software telescope") use case and resource needs - 400PB/a
DUNE scales w/ 150EB/year unrealistic. Research goes into reducing that rate. Will end up in a range between 4 and 60 PB.
LSST about 50 PB/year
Slide 17 comparing all the data scales
Future processing needs not given; current usage on slide 19; other communities benefit from our resources
Large, even daunting challenges.

Q by Mike Sokoloff: data projections look like problems w/o solutions; may need more resources
A: that's what the workshop is about; need as well to make more out of current resources. There is a shortfall with the current model; may need to get rid of intermediate data storage/steps; need to seriously rethink the model for data in particular (e.g., no room for 300kB AOD or even 30kB miniAOD)

IHEP Experiments and Software Development (Xiaomei Zhang)

Summary of all experiments going on at IHEP; DayaBay, JUNO, LHAASO, Ali, HXMT/eXTP plus their timescales.
Resources at IHEP - 20k cores; 15PB data; factor 10 in ten years planned
Network being worked on to improve international cooperation; joined LHCOne. 10 Gbps to each US and Europe
General challenges are increasing complexity in #experiments, their complexity, and resource needs; Manpower not scaling sufficiently. Thus looking for synergies and simplification.
Scientific details of JUNO experiment; optical photons dominate background correlation between events important
On software using parallelized Geant4 now; optical photons dominate simulation time; move to GPU gives x1000. Machine learning another approach; reference to two papers given

Reco may use neural networks
Analysis will become significant part of BESIII data processing; I/O could become a bottleneck
Gaudi dropped in favour of custom framework SNiPER; using TBB for parallelization now; heading for using SNiPER as common framework at the lab
HTCondor and Lustre as base for HTC computing; HPC farm being built up now; how to interact properly?
Distributed computing based on DIRAC. Joined community in 2016
Conditions databases, bookkeeping can also be areas of cooperation.
Maintenance with larger scale becoming a problem; HPC and HTC integration not easy
Cooperation with other communities via various workshops
In summary - challenges from neutrino experiments; hoping for more cooperation with other experiments

Q to slide 8: S&C work is not well recognized; do you think the situation is changing? How to get more recognition?
A: serious situation; even more in China. Much smaller community and less people;

Q: recognized a lot of the pain; best improvements were high availability dcache. Hoping for other software; downtime is painful

CWP Roadmap (Eduardo Rodrigues)

Showing timelines of various communities; operate on similar time scales for upgrade and new efforts
HL-LHC as a huge challenge for disk and computing and software
SW&C challenges far from being solved; resource estimates as in Liz' presentation
Comparison of technology evolution vs. HEP code progress
Scope and goals of the HSF
CWP process and timeline; link to final result and supporting material
Going through main points - generators and the challenging interaction with the theory community; resource needs will grow significantly because of NNLO and beyond
Detector simulation important topic and a dominating resource user; forward to dedicated session
SW Trigger and reco - move more of trigger into offline and software. Forward to dedicated session
Data Analysis - important is "time to insight". Need to be more interactive
Data processing frameworks - a lot of frameworks on the market; need to update them and consolidate
Machine learning - not a challenge, but an opportunity; need to strengthen links with other communities
Conditions data - size not a problem; opportunity to upgrade and increase commonality
Visualization - new technologies around; update and consolidate; take data formats into account if one wants to have commonality
Data management, organisation and access - storage costs are a major driver for LHC costs; lots of R&D topics to tackle; look at solutions outside community again

Facilities and Distributed Computing - need to understand effective cost; links to other big data sciences
Security - large infrastructure needs security efforts which are not ad hoc. Cooperation again
Data/Software/Analysis Preservation - challenge is to preserve bits and knowledge; analysis preservation portal of CERN is a good basis for future work
Software development - experiments have been very active recently; git, gitlab, CI, CMake etc have been huge improvements; more share means more share in training possible
Training and career - need to raise expertise; various levels of knowledge involved; how to get the proper level of training for the different people? Recognition is important!
S&C for Big data journal is a good place to make things citable; zenodo another one

Master direction is to avoid HEP-specific solutions; we are not unique in our problems any more
CWP process concluded and was a success; now need to take it as starting point for future activities
Advancing from here - need investment; there are existing efforts; diana-hep for example; CERN's R&D programme
"C-W-P" - personal interpretation; can we project? Community work packages; community-wide projects
Propose → R&D (work+ deployment)→ conclude

Q postponed to general discussion session

US Software and Computing - R&D Projects (Peter Elmer)

Snapshot of what is out there. Challenges of next 10 years repeated.
Long-running efforts as baseline and operations (e.g. US-ATLAS / -CMS, University research programs)
Additional R&D efforts. Showing CWP-relevant projects.
Going through simulation, reconstruction, in particular tracking efforts
Multi-threaded frameworks
DIANA/HEP for analysis
"Big data" projects looking at e.g. spark
List of HEP data analytics projects
HPC being pushed by politics; how to deploy software and how to interface to outside; new architectures a challenge
SLATE as a project for deployment of higher-level-services in a standardized way.
Rucio as candidate for more common data management tool
DOMA - various prototypes around. ROOT-aware SW defined storage w/ column wide data storage; Google data ocean
Training - reiterating the need. US has no HEP school yet. Working on that
S2I2 conceptualization - 2 year NSF funded project; created roadmap and plans. Ball in the court of NSF again

Q: for workload management systems - HEPcloud is not the same as BigPanda. Shouldn't go to the same slide. A: agreed.

Comment by Mike: DOE wants the HPC centres to do large part of HEP computing; HTC to HPC mapping does not seem easy; need to make sure that HPC future maps better to our needs

Q: is there inter-agency collaboration? E.g. NASA;
A by Miron L : academy of science report for NSF future plans - to support both workflows ; increased pressure for HTC in other communities; key to scientific discovery in many fields; US has investment not covered by Pete described; not visible to this community;

Liz: A by Nasa involved in LSST a bit. Only small parts at the boundary.
Ian Bird: this is not specificly a US problem - general that does not need to US specific solution
Liz: Solution is to go via international collaborations/experiments; approached by swiss centre to use more of the HPC centre. In Europe things seem more integrated between HPC/HTC
Ian: it is still the same problem; machines are not built for our problems; need to solve it once and for all
Liz: HPC like “snowflakes” - All are different.
Swiss request (CSCS) was driven by ATLAS success. 5 supercomputers joined (Phoenix?). And another one is interactive.
Peter: who In Europe/HPC is enabled to push into a different direction. Ian; the experiments; through praise and HW being provided; opening up opportunities w/ praise and influence, Open science cloud of EU an opportunity.
C by ?: to be the biggest user (lattice QCD e.g.) is helpful. When we become the big users we will be sitting in the right room with the people. Have to be involved with every system. Every single machine is an independent investment.
C by Mattias: strong case for influencing the science; having new middleware to abstract away; throw away the complicated 10 %
For HPC systems it is important were to be in TOP500 and not what science is done there. HTC seems not compatible with this. Motivation is HPC, not HTC. otherwise money would be spent in a different way.
Graeme cutting off the discussion on HTC/HPC. Challenge would be first to make our SW use our current resources used efficiently.

Discussion Session

Presentation by Graeme.
What is the best way to continue now? What do people expect? What do they want to give? How to motivate people to give extra effort beyond day to day work? New projects? How to encourage more explorative solutions? CVMFS is a good example for success there.
By-product of another R&D activtiy

Comment by Paolo: HSF/CWP was a nice dream list. Now the question is - can the HSF become a place for a program manager to invest?
Peter: they put money into something. Is that impactful? Want to maximize that. CWP mentions important pieces. Question is what makes projects successful? Phedex and Rucio

as two examples - why do we have two projects? Historical reasons? How do you build a community?
Q by Jan Iven: can HSF give a recommendation to funding agencies about which projects are worth funding? How to select from competing projects?
Peter: actually not all projects are visible even.
Dave: one of the problems is the lack of scale? How many FTE are actually needed? What will be the share of the individual funding agencies. One has to set a scale for the funding needed.
Ian Bird: one is missing there is the WLCG strategy view on this. LHCC waiting for it. What are our (WLCG) priorities. Identify high-priority; sketch out and estimate resources; hope end of this week to make draft more widely available; could be a base for talking to funding agencies;
HSF is not top-down; WLCG is;
Pere: problem was giving global priorities in CWP; Liz: that was not the intention. Wanted to be inclusive and get everybody involved. A top-down entity is needed though for the next steps.WLCG is one of the possible ones. If we want to go where we need to be, we need a world-wide computing infrastructure.
Ian: proposed multiple times; people don't know how it should look like. WLCG as an example, but w/o LHC focus. And then on top the experiment specific resources. Need that for LHC, neutrinos, etc independently. A path from current situation to that not clear
Peter: <can somebody in front hear him?>
Graeme: contributions to common efforts. Important that not everybody takes the most sexy projects.
Ian: separate two pieces; you buy resources for ATLAS; operation and middle-ware shared by everybody. Justification different
Paolo: WLCG strategy document needed. But is that good for research programs? Difference between operations and research program.

Graeme: another comment. How can HSF help? We had two community reviews in the very open. Gather experts and identify their needs. Not only to cut off projects, but to guide them for doing the very useful things
Oli: good project based on good stakeholders. HSF can play that role. Comparison with Apache foundation. By conducting community reviews by experts to answer questions like what sw should be retired and what should be elevated.
Brian: have a good eco-system for projects. And then see how projects evolve and experiments can use.
Pete: <still not understandable in the back>
Mike: HSF is a community organization w/o resources. CWP is a way to communicate what we consider important. WLCG is different. For program managers is *a* voice of the community. Can summarize the person-power needs w/o assigning priorties.

Technology Watch

Evolution of technology and markets

Presentation by Bernd.
Recent RAM price increase (120%) -> revenue growth. 3.7T$.
Expected increase in memory 20% next year. Smartphones saturated, market is now replacement. Current processor tech can go to 7nm, after which tech investment is large.
Many new processor archs coming onstream soon - driven by machine learning. Local data processing becoming more important with these archs. GPU market very large driven by gaming and mining. Other accels unclear (e.g. end of line of Xeon Phi).
FPGA devices becoming more used. 50-70 qubit quantum computers, and fabulously expensive - not clear how LHC could make use of these.
DRAM has gone up 150% in test 24 months, still 20% to go, and then new fabs should help control the cost increase. DRAM scaling slowing down now.
Various new memory technologies on the horizon, but still not cost effective.
NAND storage price trend not clear. Capacity growth is now linear.
HDD price/space evolution flattening. 14TB capacity with He fill, prob get to 20TB. Two new approaches for higher density (Seagate prod 2020 vs WD prod 2019) - 20/25/40TB per disk. Multiple independent heads - keep IOPS/TB constant.
SSD not likely to replace HDD any time soon for us.
Tape LTO-8 (12 TB), poor market now, only one supplier, market shrinking- need to watch market.
Servers market now 21B$, pushed on by hyperscale data centres (google, aws, et al)
Intel dominating, 99% of market. Possible contender is IBM Power 8, AMD EPYC, ShenWei (260-core), ARM (power/specint interesting, price still x2).
Service costing evolution - recent kink due to memory price hike.
Assume: 15% price/perf improvements for server, 20% for diskserver.

Summary: price/perf slowing down. New processors driven by ML. HDD still the thing.

Paolo: in your plot, considering peak or actual performance?
B: this is HepSpec.

?: evolution of tape? Will disk become cheaper than tape?
B: never, more likely risk is that tape will disappear.

What are we doing wrong - why facebook not using tape?
B: access patterns are different, though they are looking at backup on tape. Also they have lots of cash for disk!

Liz: Google was trying to convince tape is finished - just turn off the spinning disks and keep ‘em cold? B: not so easy, electricity costs are not always the factor. Ops overhead, disks also age. Makes erasure coding a challenge.

Brian: when will other academic communities star to affect the price point. B: not at all, we're too small. Google / banking / etc dominate the customer market.

Graeme: processor market - is that one of the healthier areas? New processors are niche. B: not a good trend for us. ML is fuzzy, so chips specialised in it are less obviously useful for deterministic science.

GÉANT Data Transfer Node service

Vincenzo Capone, GEANT.

Problem: TCP is rubbish on standard servers, compared to the bandwidth we could make us of. Packet loss on WAN kills TCP, so network is engineered for 0 packet loss as 80% of transfers use TCP. SO we use optimised tools (gridFTP/FDT).
DTN service: dedicated hardware for data movement, carefully place wrt network, no firewall. Used for troubleshooting and validation, to understand where the network problem is. Tests end-to-end storage/network/storage stack.
DTN machines already on GEANT bakbone (London-Paris).
Would like to understand things we might need to make the service useful for us - contact us if you'd like to participate.

Oliver: DTN not just at network level - it uses a new data endpoint? How does it improve the performance. V: It uses your standard data storage software to test your whole stack including the network.

Alessandra: compared to perfsonar? V: not alternative to perfsonar which looks in much more detail at network performance.

Topics for R&D on technology watch

Bernd.

Many things we could look at.
Hardware: next gen processors, next HDD sizes, server form factor, nextgen switcher/routers
Software: config mgmt, containers, VMs, SDN.
Is there are need for more coordination of tech watch?
"Everything is connected" - hard to determine what is"better".
refers reader to slides
Expensive to evaluate all of this.

- How to investigate the side-effects of what we decide? I.e. TCO.
- Forum to discuss

Topics where we as a community need to spend effort to watch what is happening?

Oli: Folds into computing models of the experiments - can use them to determine the impact of these things. Site implementation is an important factor and can vary a lot.

Jan: how to coordinate the site and the experiments?

Miron: question: openlab with intel, looking at different tools to partition the resources between different jobs running on the same server. Do we care about how memory / cache is being split, to maximise the efficiency? Graeme: we don't how much we care until we know how much of an impact it has. How to change software to make use of it? Miron: look at profiling the application, how much effort are the experiments able to put into looking at these concepts?
Marco: yes, need a collaboration, but unlikely to study these in detail on applications running on a T2. HLT yes.

Proposal for technology watch WG

Helge.
Proposal for WG to do the tech watch.
Becoming more important as we go to future runs, better to be shared over more people.
Start WG, open to all interested sites and experiments.

- Usual reports to WLCG/HSF, 'EPIx, CHEP and GDB
- Written reports? Living doc on web?
- Live forever, but review every year

Miron: assuming clouds offer some value, what should we rent from clouds? Becomes difficult to decide what to focus on in a fast evolving technology landscape (e.g. GPUs).
Helge: yes, need to track, but expect the answer to change with time.
Bernd: requires close cooperation with procurement of every site
Ian: one fo the uses of this tech watch is to explain to our funders how well we use the money. So yes, shoud compare what we are doing wrt commercial cloud offerings.
?: internal costs sometimes rather murky and hard to determine in the academic community. Hard to work out.
Bernd: has a comparison (for CERN) already, but this needs to be re-done continuously.
?: 50% difference with cloud is rather optimistic. Liz: spot market varies a lot, not clear what the real number is. Ian: FNAL comparison was all about the hardware.

Miron: people will keep some money for elasticity - this effort should aim to help how best to direct that money.

Ian: expand "tech watch" to "cost watch" since some of the options are things like external cloud or HPC, that we don't get to control.

We need to be sensitive to the developer and operational costs of running on different types of hardware.

Liz: fully support, and needs basket of applications to determine how well the things we actually use performs on these new things compared, say to just using HS06.

Brian: don't spend all the time pimping the code for specific archs, rather than concentrating on general code quality.
?: yes, agreed that we should focus on good code.

Ian B: sounds like we should **merge this will the Benchmarking WG**.

- Jan: Some dissent - keep groups separate, but exchange information as needed.

NDAs? Maybe just better to read The Register!
Not clear focussing on NDA'd technology is a good idea, since it's hard to have an open discussion of it.

General agreement, but proposal needs some refinement about potential merging with BenchWG.

Helge points out the existence of the WG "Systems Performance and cost modeling" working group ; the "technology watch" group and the "cost modeling" should work in close collaboration.

Sponsor presentation: DELL

Sponsor presentation: RITTAL

Use Cases

Introduction

Develop from the challenges presented in the morning. See later sessions for many more details on other topics.

Challenges for LHCb trigger

Upgrade will allow higher pileup running. Very high signal rates for the LHCb physics programme. Readout detector at 30MHz BC rate from LHC. Implement a software trigger and do analysis directly from DAQ.

This is an extension of many pieces already in place for Run 2. Synchronous HLT1, asynchronous HLT2, TURBO stream. Need to perform offline quality reco in the HLT farm. 2-5GB/s to storage. HLT1 reduces rate to 0.5-1.0MHz. HLT2 gives best tracking performance.
Gap between GFLOP rate in farm and number of trigger decisions taken is growing (c.f. Effective FLOP usage this morning).
Gaudi needs to be modernised and make it fit for Run 3. Memory savings are impressive using MT. Modernising code to new C++ standards, improving the code itself - helps with runtime performance. Vectorisation required a refactored data model, but brings real gains (Kalman filter).
VELO track reconstruction optimised to give x3 sleep-up.
HLT1 now runs faster in multi-threaded mode than multi-process (+20%). Detailed improvements, algorithm by algorithm.
50% of resources are spent on ‘data preparation’, as opposed to number crunching.
Interplay between physics study and cuts is important - some studies suffer much wors, others better.
Not yet reached 30MHz, but well on the way - still 2 years to improve.
A lot of effort in hackathons and training - improve literacy and effectiveness.

LHCb planning upgrade II at L=2.10^34. All BC have signal, but a lot of pile-up contamination. Naive scaling would be x10 over Run 3 - this becomes a GPD level of data taking and processing.

Q. Did you measure thread-contention in MT? A. Yes, it’s been measured, many points were improved on to get to the results presented today. Currently using an event level parallelisation approach (not task/alg parallelisation).

Q. What about GPUs? Did you look to use them for reco algorithms. A. Yes, we are looking at them, but no results yet. Might see use as a ‘pre-processor’? TBD.

Simulation and event reconstruction challenges for DUNE

Review of physics of neutrino mixing. Big questions: mass ordering? Is their CP violation in neutrinos? More than 3 flavours? Is general neutrino picture correct?

LNBF facility studies this - LAr TPCs are the detector technology (MicroBooNE, DUNE). Single phase detectors are simpler. Dual phase more sensitive.

LArSoft is common code shared by all the LAr experiments. Significant fraction is shared.

2D view matching is used - clusters. 3D doesn’t cluster, go directly to the reconstructed event. CNNs very interesting - pixel by pixel classifications. Hybrid approach combines with domain knowledge. Different experiments combine approaches

MicroBooNE very good energy resolution, but shower algorithms can under cluster. Being improved. Track/shower separation difficult at low energies. Short tracks are problematic. Event vertxt sometimes goes wrong.

Improved algorithms take more time!

Simulation is …

Scale issues: DUNE far detector is very large and very high resolution. Could have 400PB of data for beams. SN and proton decay would increase that. Need to manage data by triggering and zero suppressing. Sweet spot to avoid sacrificing physics, but manage data rates.

Simulation and reco are not fast. Minutes per event. No hot-spots are seen. Simulation can be 6-8GB. Reco > 2GB. Multi-threading. No use of SIMD right now. Investing in software optimisation is a big focus of effort. See a lot of promise for ML - already a lot of good results. Can CNNs be used on the grid. Convention algorithms will, however, remain an important component.

ML running on GPUs on HPCs - is this an option?

Q. What sort of pre-processing is needed for the CNNs? A. These are applied at pixel level. The inference is very expensive - can take minutes. Other techniques work better for different algorithms.

Q. How do you manage to share code? A. Based on the fact that the data does look very similar from these detectors. Community is quite close and that helps.

Reconstruction challenges for HL-LHC

“The times they are a changing”.
Pileup 200 increases occupancy by x4. But we want to do a more ambitious physics programme. New detectors will be installed with higher granularity. Timing information will be there. This makes tracking easier. GPD 1st level trigger will go to 1MHz (x10). Final output rate will be 10kHz. Conclusion -> we would need x100 more resources.
Pileup could mostly be identified in the tracker - just ignore it in reconstruction.
Dream would be to identify and measure all particles in a collision.
Technology improvements are not coming so fast these days.
Memory latency is becoming a real issue - much longer to fetch/store than to calculate (FMA). Theoretical throughput cannot be supported by memory bandwidth. Can sustain only with ~100 computations per memory load. Thus, need to do *more* per piece of data, not less.
Data reduction is really important - 1kB per event is possible. But as this is just 1bit per particle, is the whole reco chain just unnecessary/wrong.

Q. Is it data cache misses that’s the issue or instruction caches? Both are important - would need 1 bit per algorithm! A. CMS never saw instruction cache misses.

Multi-threading multi-event model identified in 2007 - takes 10 years to move key ideas into production. Don't prematurely optimise. Our software has a very long lifetime. I/O is a serious problem, but significant improvements made (even histograms can be made to work). Multiple algorithms in flight can help reduce memory needs and improves locality. Parallelisation requires SIMD/SIMT in the innermost loops. Branching breaks this very badly. Adapted algorithms become very fragile. Need new algorithmic approaches - keep data local and enable concurrency. Need to care about thread ordering. Do not transfer data needlessly. Simple data formats are really best.
-> Accelerated application that takes RAW data and produces physics.
Porting clients to data structures takes a long time and is very tedious.
Different models of GPU *integration* are possible, but actual deployment is a very long process. Integration of COTS machine learning oriented hardware will be difficult.
New paradigms: artificial intelligence, specialised hardware, smart data mining?
Really need new R&D for new challenges with few people.

Q. Do we need to do this multiple times for each experiment? We need to do more to abstract details and share much more in terms of software. A. But how to deal with real differences between detectors (e.g., stereo strips)?

Challenges and evolution in event generators

Differences between ATLAS and CMS generators consumption not well understood. Generators do a lot - hard scatter to hadronisation. LO generators not too hard. NLO widely used, complexity grows with factorial of particles and loops (e.g., 100k diagrams for W+5 jets). Memory consumption becomes a problem too - need to store in memory. NNLO is needed as some experimental errors are now < NLO errors. NNLO limited to 2 jets at the moment. HPCs are being used to speed up time to results (doesn't help with size of computing problem itself). Work with generator authors to increase performance. One concrete issue is over object-orientation - too many function calls with small code blocks before exit (Sherpa). MadGraph would need to rewrite their integration step to parallelise beyond one node. New integration techniques being looked at - new MC techniques or ML driven phase space integration (SciDAQ project).

Q. CMS plots don't have the phase space integration - so why so different? A. ATLAS uses Alpgen and GENIE that are more expensive than MadGraph used by CMS.

Q. Do generators run well on HPCs? A. Yes, overhead from orchestration is very small.

Q. For NNLO what is the resource usage? A. Girdpack can take a week on a single machine. 25 million hours per paper from Mira run. Some events can take 10-15 minutes on Sherpa - when it needs to flip to quad precision because of a cancellation.

Q. Can the generator community benefit from modern coding styles? A. Sherpa have 27 layers of virtual inheritance - that probably could be much improved with a rewrite.

Tuesday

Common Data Management and Data Lakes

Introduction

Highlights of the Rucio workshop

Q: what's the development community, who's involved?
A: CERN, Oslo, Wuppertal, Innsbruck, Argentina + other contributions. Plan up to 2025 targeting hi-lumi, inc data lake, caching, sub-file tracking.

Q: Happy to see HTCondor integration. What's the integration? Maybe "interface" is better.
A: Agreed

Q: Google cloud - does that use FTS and third party copy
A: Streaming for now, 3pc when support arrives in davix

Q: Can you publish datasets with digital identifiers (OID)s?
A: The data model supports it, but no integration at the moment.

Dynamo - Intelligent Data Management

Q: Session title is "common data management" - why did you build this when Rucio is already there. How does this fit into the "common systems" philosophy?
A: Developed 1 or 2 yrs ago. Would be good to have a second experiment testing the software. When we have docs ready. There are no plans at the moment, have to see if we make this available, or if this is the final choice for CMS

Q: With inventory in memory, what's the footprint for, say, 10M objects?
A: Current system uses 10GB RAM. Don't foresee scalability problems.

Q: How do you manage data transfers? Which protocols and which endpoints?
A: Info from all sites on protocols etc. Tape - not online at the moment.

Q: Atlas discussed all this in Naples 7-8 years ago. Started with popularity. How do you place data based on popularity?
A: We have a queue of requests with a ranking by popularity, which we process from top to bottom, destination chosen randomly. Trying to quantify how this benefits dataset availability

Data Lakes R&D: high level goals

Q: How did contributing sites give their hardware? Sites typically support lots of communities, how does that fit?
A: Variety of approaches. At Dubna we got VMs and used a container. SARA was rpms. RAL VMs which have Ceph. We accept anything. We imagine a data lake as a multi-community solution.

Q: How do central operations workflows have to adapt to this (e.g. in Atlas or CMS)
A: We're moving from one distributed model to another
A: We have to evaluate this new way of managing the existing infrastructure to see if it's advantageous
Q: We don't know much about requirements or impact, but it seems the answer is EOS. This is dangerous and we should split these, get the requirements straight first to make the answer credible.
A: The answer is not EOS, this is just the system we're using to test. Ceph and dCache also have advantages.
Q: In 3 weeks you could also build a distributed dCache like in nordugrid. Be careful how this looks from outside the community.
A: Part of the answer to this will come from tomorrow's discussions and show the potential of a particular prototype.

Q: slide 5. We see CPUs on the slide, so full representation of existing stuff. Are we talking about a single new service or an ecosystem?
A: Could be either. Experiment and explore.

Q: This is an evolution of the existing system. How do you evaluate the cost of network cf storage?
A: Agree this is an important question.

A Transfer Ecosystem for the WLCG

Q: This is https based. Can you connect this to http federations and the data lake?
A: This is for bulk movement, rather than getting data to clients. CMS uses data federations and focuses on subfile reads with intelligence in the client. Not sure how close these things will get.
A: Dynafed team at least is ensuring that these things will work together. XDC is also looking at this.

Q: WLCG cares about audit and who did what, how doe this map
A: This is discussed in the auth TF, there's difference between traceability and authorisation. The tokens give traceability, you can see which token was issued to whom. Alice has shown this can be decouples from authorisation.

Q: Can you talk about interoperability between different authentication systems?

A: We're working on a common profile for the tokens. Differentiate identity tokens (who you are) and auth tokens (what you're allowed to do). See Maarten Litmaath's talk.

Q: X509 scales very well. Bearer tokens need a central service to create them. How does this scale?
A: We use distributed verification so you don't need a central service for that, more details tomorrow.

Performance metrics and measurements in the Data Lake model

Q: Still a bit confused about data lakes. I thought these were a model of serving smaller tier sites from larger tier sites. Instead we are looking at another data management layer. How are data lake developments linked to things like rucio and dynamo.
A: This is an R&D activity about understanding how distributed storage would work + ecosystem. This ecosystem includes dynamo, rucio, caching, transfers etc.

Q: LHCb will be the same size as Atlas and CMS in terms of data management. We are writing a computing model for R3 this year. We need to know how data lakes affect this model. Why aren't we involved in this study?
A: You were invited. However, we need to include LHCb use cases.

Q: Wld you like to include NDGF T1 in your prototype infrastructure? We've been doing this for 10 years.
A: yes.
A: We will discuss interop between dCache and EOS in the dedicated session.

Wrap Up

Frameworks and Infrastructure

Threads to grids: distributed concurrency in athena/Gaudi

Motivation: HPC needed to close resource gap

- Next-gen HPCs will be co-processor based, with GPGPU certan to play

Issues:

- Each HPC system is unique. Porting cost non-trivial. Optimization cost even higher.
- ATLAS and CMS developing/prototyping "GPU services".
    - Buffering data in mini-batches from multiple events will be required
    - Considering shipping input data from multiple nodes to GPU node to keep GPU fed (ATLAS EventService on steroids)
    - Can we turn access to event store into MPI-like messaging?
    - Can we get all data needed (event and non-event) into a User Algorithm via transient messages?

Common Developments:

- Heterogeneous scheduling simulator to study event throughput as function of multiple processing parameters on different simulated architectures

Common Issues:

- Offloading services
- Algorithmic code conversion
- Efficient marshalling / unmarshalling of data for interprocess communication.

Q&A:

Q: We should not tie this R&D to supercomputers. We need a more abstract way of writing computation kernels. Like the idea of a simulation to model this without porting everything. Not an option to do nothing given huge HL-LHC requirements.

Q: GPUs may come and go, but computation intensity will stay, and we need to improve that metric whichever architecture we target.

Q: We should not focus on supercomputers, but we should rethink the processing paradigm, possibly splitting events so to process many of them at the same time.

CLARA: Clas12 Reconstruction and Analysis Framework

Motivation:

- Data volumes of 2020s experiments like SKA HL-LHC
- HEP monolithic C++ code is not encouraging new generations to join the field.

Approach:

- CLARA based on Micro-services glued by flow-based programming.
- Message-passing data communication.
- Multithreading for vertical scaling at the service or service method level
- Horizontal scaling on cluster relies on network communication (ZeroMQ)
- Native multi-language support.
- Algorithm composed by graph of mu-services
- Reactive data-driven stream processing.
    - Can be self-routing with no overall scheduler
    - Control flow supported based on mu-services state.
- Rich user interface allows local or cloud deployment provides monitoring information.

Wide variety of applications to multiple experiments and groups including theorists at JLAB. Interest from NASA SRB group.

Q. What's the penalty for moving data? A. Yes, it does cost, but there are ways to minimise it (run 'close'). But processes running on the same machine use shared memory, so nothing needs to be copied.

Q: What are the containers you mentioned?
A: It's a logical grouping of microservices to be used as an *application*.

ALFA: Alice-Fair message queue based reconstruction and analysis framework

Motivation:

- FAIR facility under construction: Few large experiment (CBM, Panda, R3B), many small experiment within APPA (O (10) people) with similar requirement for simulation
- FairRoot developed over last 15 years originally for FAIR experiment and later for others (NICA project in Dubna)
- Similar requirement in ALICE Run3 (>1 TByte/s) and the large participation of the GSI in ALICE experiment lead to collaboration on software ALFA

Approach:

- FairMQ data processing layer based on multi-processing with message queues introduced in ALFA.
- FairMQ Devices has Data Channels (message passing) and call backs to the user Algorithms.
- Several messaging and serialization protocols supported by FairMQ transport layer.
- Each Device is a separated process, that
    - can be multithreaded.
    - can run on different hardware,
    - can use different languages.
- User only sees data channels and focus on algorithmic code.
- Like CLARA, FairMQ uses micro-services under the hood.
- Based on industry standard technologies e.g. for message passing.
- Dynamic Deployment System deploys application on e.g. cluster based on topology description in XML.

Examples:

- ALICE parallel simulation will be described in other talk in this workshop.
- PANDA-MVD DAQ.
- ALICE Run 3 GPU Tracking. GPU TPC tracking in operation for 10 years. Significant experience on effort needed for migration. Worth gains. 20GB data frames O(100K) tracks each.

Q: Is just a small team of people working on this or you need training a for a large group?
A: Need some specialists as some of the tasks are not really in the profile of a typical physicist.

Q: How do you manage resources with all these small independent services?
A: It's responsibility of the developers to declare the resources needed by the device, then DDS is using watchdogs to make sure that the requirements are under control.

Q: DDS has the same role of CLARA's orchestrator, how do you handle exceptions?
A: Message passing is using push/pull to make sure that messages are correctly delivered. For devices, we know that they may fail and we let them fail.

Q: Did you measure performance? On slide 27 there are only 2 *important* tasks, all the rest is *overhead*.
A: We measured compute/message for some examples. The important point is that the ratio is such that the overhead of messages is minimal.

Ideas for Framework Discussion

- Common scheduling simulator (building on MT scheduling simulators. Include dummy algorithms, data transfer bandwidth, co-processor latency, network latency, etc
- Common interprocess/internode communication mechanism. Include zero-copying buffer transfers, shared memory, serialization/deserialization, network/pipe transfers
-

Training and Careers (Parallel)

Introduction

Points of discussion for the panel:

- What has been accomplished in software training?
- Are we meeting our goals already?
- What's missing? *i.e.* we have lots of bottom-up initiatives but lack of coordination
- Other approaches that can save money and increase participation (web-based self-training, local schools…)
- What's a good balance between generic/specific training? *i.e.* programming languages vs. experiment frameworks
- How do we pick our teachers? And how do we make teaching appealing for one's career?
- How do we favor collaboration between existing schools/efforts?
- How do we finance schools?
- Propose training as a service task?
- Have an official HEP software curriculum so that teaching activities can be more easily accepted and recognized by the experiments

Lightning Talk 1 (CMS/ATLAS)

CMS has a Training Model which is not clearly recognized (it's mostly users self-organizing, no dedicated coordination role). However, the training program is robust, and engaging in

training activities gives you recognized credits. CMS has a long tradition in training (started in 2006, before taking data): classes of different lengths (one-day hands-on to one-week-long schools). Teachers are recognized experts in the taught subjects. CMS teaching material is available on TWikis and Indico. Some CMS schools favor gathering in groups to achieve a goal, which is eventually evaluated and possibly given prizes. Not all CMS schools allow for remote participation via Vidyo, but some do. CMS has a Career Committee that audits people's careers after quitting research: HEP computing skills are appreciated in industry.

ATLAS has most of its material in Twikis and talks on Indico: currently very difficult finding people to maintain, and **not a single place where material is collected**. ATLAS lacks intermediate level training: only very basic (ROOT…) or advanced.

Discussion: is performance optimisation the main goal? No - we have a lot of important training to do and it combines (physics, performance, sustainability).

Lightning Talk 2 (Belle II)

Belle II has custom software for reconstruction/analysis: this requires specialized training. Among the user needs, there's generic tools (Python 3, LaTeX…) and specific ones (statistical software…) Non-software-related concepts such as flavour physics/statistics are required. Training material is available in wikis and software manuals (using standard formats like Sphinx). Manuals are less likely to get out of date as doc goes with the code, whereas wikis get outdated quickly. Two support channels: forums (Stack Overflow-like), and chats. Missing a single point of entry. The Belle II "starter kit" is made of Jupyter notebooks on a dedicated server: **doc + code go together** and it's easily portable to other institutes. First "starter kit" experience: very well received. Manuals are being migrated entirely on Sphinx. Challenge: provide pupils with skills easily reusable in industry (most of them will leave HEP). Addressed by teaching mainstream tools as much as possible.

Discussion

Q: How do you assess students' progress?
A: One of the metrics is opening a chat room during the tutorial. If the chat room gets very busy, it probably means what was explained was unclear.

Lightning Talk 3 (The CERN School of Computing)

CSC focuses on generic "knowledge" rather than "know-how". Works as a University: has a final exam and gives ECTS credits. Apart from the main "generic" CSC, there's the thematic one and an "inverted" one every year. Basic CSC: base technologies, physics computing, data technologies. Thematic schools focus for one week on a specific topic (e.g. parallel computing, optimizations…). Inverted schools have selected lecturers who were CSC students. Being two weeks long, CSC students are usually lodged in a "campus" and a social/sport programme is offered as well, in order to ease networking. Material is available as a printed booklet, and exercises are also done in pairs. Exercises are done on VMs. There are limited places, and students undergo a selection procedure (not

first-come-first-served). Student feedback is always very positive. Students are a mix of physicists and computing engineers. Difficult not to teach specific technologies and keep being “generic”. It’s a challenge not to teach hype topics. More diversity is also required (way more men than women).

Lightning Talk 4 (Starterkit - LHCb)

Why the starter kit: students are trained as physicists but asked to be data analysts. The goal is to make the capable of starting right away their work. Tutorials are based on gitbook, contributions via pull requests. Workshops have a small fee (25 CHF), pupils split into small groups. No Vidyo (require presence). Three levels: basic tools (non-HEP) such ash bash, git, python; basic experiment-specific training; advanced topics too (“ImpactKit”). StarterKit has 40 participants, 4 days long - ImpactKit has only 20 participants, and ends with a hackaton. Once per year. Fully managed by students, continuous rotation - ensure knowledge transfer. Method is very efficient and appreciated: today’s pupils are ready to be tomorrow’s teachers. However, lots of organization required: room booking, too many volunteers needed, coming to CERN not convenient for everybody. 2015: first effort. Last year (2017): joint forces with ALICE, very positive collaboration with a common repository for basic courses. StarterKit pages are true documentation in LHCb. Feedback is positive: people feel supported by teachers, even if they think teachers go too fast on some LHCb-specific parts. Documentation is reviewed every year before the workshop (always by new people). Problem with the model: does not scale a lot. Suggestion, maybe? Take the StarterKit model and organize your own! Note: the social event is very important, as well as coffee :-)

Lightning Talk 5 (ALICE)

Training activities are improving since 2014, positive experience. Focused on enabling analysis activities for attendants. Tutorials online (git)
Workshops organised around thematic topics (1-2 hours) with online connection + basic software (attended by newcomers). Teachers are experts in the topic.
ALICE + LHCb joined in beginner topics on general computing, mixing teachers and adding Vidyo and mattermost channel.
Wide variety of topics, adapt to also time-sensitive topics (such as move to git pull requests)
Growing interest in python, even if officially only C++ is supported
Feedback: lessons should be longer, users appreciate conjoint efforts with LHCb, also coming in person is seen as good (networking, meet the experts), events can be used by the teachers as Q&A for developers.
Open points: difficulty finding people to help (not rewarding career-wise, time-consuming), so tutorials end up associated to certain people, how to vet and teach teachers?, how to involve people far away from CERN? Should we organise TED-x like training sessions?

Lightning Talk 6 (Data Science @Università di Padova and INFN)

MSc degree -> Physics on Data (purely academic) starting in October. Merge competences of physicists and data scientists (meet job market requirements + meet research requirements).

Built on a commentary and editorial on Nature:

- Improve exposure of students to stats and probability
- Improve knowledge of mathematical tools, such as machine learning.
- Train a new generation of physicists mastering these tools

International, taught in English, limited to 40 students, need sufficient Physics knowledge, mandatory internship in company.
Mix of mandatory and optional classes, possibility of specialization (Fundamental interactions, physics of matter, physics of the universe). Both tech classes (ie machine learning) and practical ones.
Pillars:

- Lab of Computational Physics, from python ecosystem (no C++) to MC methods, and the moving on the machine learning. Large computing infrastructure.
- Management and data flow of datasets.
- Applied statistics

Getting started

Lightning Talk 7 (Bertinoro School of Computing)

Motivation: good training material and schools help the real experts concentrate more on Physics and less on debugging bad software written by untrained people. School teaches modern C++, memory-friendly programming and efficiency (vectorization, parallelization, GP-GPUs…) School is quite dense (5 full days, 90 minute sessions). Interaction between students and teachers is paramount: this is why coffee breaks, lunches are long, as well as the “consolidation time hands-on” at the end of each topic. Challenges: keeping constantly updated with new technologies, do not overload students (school is very dense…), find a common thread across topics. Feedback shows an improvement over the years. What is required to be a trainer? Need to learn what you teach, be confident with your material - it requires lots of preparatory time (study first; follow up on forums, mailing lists, write code…) In order to find many teachers, the school needs to motivate people to teach. Trainers need to be involved both in large and small experiments.

Lightning Talk 8 (Neutrino experiments @ FNAL)

Several communities involved, not just CMS. Neutrino/Muon communities: common sw stack (art)... unfortunately the teams maintaining the stack do doc as well, but little time for that, so: incomplete. Students are mostly on their own for programming languages: experiments provide references (to C++ guides for instance) but with no indication of what’s important. Requirement: follow some quickstart that gets you started in 1h from zero. Unfortunately, even if the stack is in common, every experiment provides its own description of it, without making a clear distinction between what’s the common stack and what are the experiment specific things. Most of the training is word-of-mouth. Slack channels are very popular. Many examples to copy from (but no vetting of the examples). Advanced training has a better organization. Particle Astro community: mostly python + native C. Each subgroup has its own analysis software. Some commonality being seeked… not achieved though. Machine learning: Deep Learning Journal Club + Machine Intelligence Group: advanced training aiming to provide workshops and authoritative support on specific advanced topics.

Unfortunately, even if startup difficulties are acknowledged, many seniors don't see training activities as beneficial.

Lightning Talk 9 (VIRGO+LIGO)

Very small communities that don't talk to each other for software. Having an existing structure making them talk to each other would be very beneficial as at the moment they don't have the resources for doing that.

Lightning Talk 10 (Vision for Future Integrated Software Training)

Three tiers of HEP software training: from the bottom up, "carpentries workshops" (git, python, ROOT, Unix…) for early Ph.D. students, then experiment training (medium level), then advanced developer training (mostly for postdocs and Ph.D.), consisting in thematic/intensive schools (GridKa, Bertinoro…). Levels provided by different institutions: if standardized, it's easier. The idea is to consolidate the tiers in order

Round Table - Discussion

(Starting from Peter's talk.)
Dario M.: Universities vs. Research Institutes, there's a gap: universities don't teach things useful in practice for starting working for a research institute. INFN (Italy) is a good example where the "horizontal" gap is bridged. Then there's a "vertical" gap, the different tiers need consolidation in order to avoid overlapping, and gaps.

In practice: universities need to be told by the "stakeholders" to teach the right and useful stuff as well for preparing software developers!

Andrea V.: start from ALICE+LHCb example: separate experiment-agnostic stuff from specific. Some advanced topics (ImpactKit, LHCb) are shown as *very* experiment specific even if they are not: one can teach C++ debugging/profiling without being necessarily tied to an experiment.

According to Vincenzo I., some topics need necessarily to be taught at universities when you are an undergraduate. StarterKit once a year is definitely not enough! Would not work for CMS.

Albert P. (LHCb): yes however StarterKit material is available online - even if it's 40 people a year attending it, everybody can read the material, so definitely more people benefitting from that.

Edward Moyse: StarterKits are the way to go. We didn't talk much about *career* however: people leave experiments, how to keep people after training them?!

Sudhir: people from the management in experiments need to recognize more the efforts of people involved in training.

People being trained in HEP experiments (ATLAS example) even if they leave HEP they mostly end up in very interesting fields.

HSF can help providing the infrastructure for making the material available online in a Coursera fashion.

Data Preservation for HEP (Parallel)

Certification: Motivation, Benefits and Status

CERN Analysis Preservation Portal

Data Access (& Discovery?) - A New Achilles' Heel in LTDP?

Collaboration Across HEP and with Other Disciplines: An EIROForum Sub-Group

Archiving and Preservation "in the Cloud" - Outlook

CWP Roadmap, Preparing Input to update of European Strategy for Particle Physics (and others elsewhere in the world)

Workload Management (Parallel)

How HL-LHC challenges inform WM R&D: ATLAS view

Have to manage expectations of analysts used to having all data on disk. Agreed; we are part way there with analysts acclimatized to waiting for the analysis derivation train workflow we have in place now, we have to make the tape based workflow comparably fast to meet expectations.

Have work to do on the whole workflow to make tape carousels work; right now we write data to tape as quickly as possible, we need to write it optimally for reading it in carousel mode. Close collaboration with data management.

How HL-LHC challenges inform WM R&D: CMS view

Same view as ATLAS on CPU vs storage, the latter is the bigger problem.

CMS succeeded in the social engineering of getting analysis groups to agree on a common analysis data product content. ATLAS failed and has ~100 of them.

Don’t need tape carousel for nanoAOD, can all fit on disk. May need it upstream for managed processing on larger formats.

In any case, worthwhile in the near term to fix the inefficiencies of working from tape.

Roadmap towards WM standardization for HL-LHC: ARC perspective

Roadmap towards WM standardization for HL-LHC: PanDA perspective

Roadmap towards WM standardization for HL-LHC: Dirac perspective

Roadmap towards WM standardization for HL-LHC: OSG perspective

Analysis Facilities and Use Cases

Overview of CWP paper from WG Data Analysis and Interpretation

Spark-like and query-like analysis systems and tools

HPC analyses

Real-time analyses - the LHCb case

Analyses with SWAN and docker

Open discussion on future avenues and collaboration

- We want to foster collaboration and common project on analysis related topics.
- Questions:
    - Do we have analysis efforts that were not covered, that should be mentioned here?
    - Is more contact information needed for projects already discussed to start more collaboration?
    - How do we keep the discussion and exchange of information alive?

Wednesday

Programming for Concurrency and Co-Processors

FPGAs as co-processors for reconstruction

Q. On slide 12 why is the single photon so slow? A. It's the cost of moving the data to the FPGA and back, one photon at a time.

Q. Are there generic libraries for ML? A. Yes, but mainly for high throughput applications with data in the main memory (good for offline); for triggers problem is more latency bound, so likely we would write our own engines.

VecCore: Expressing HEP algorithms in terms of explicit SIMD types and operations

Comment: CUDA seems to be an easier way to parallelise.

Q. Isn't it worth just supporting VC instead and not developing another 'homemade' layer, especially if we foresee that VC will evolve into the C++ standard? A. VecCore supports multiple backends (e.g. CUDA) and gives an abstraction layer that can be useful for switching backends. LHCb need to support ARM and Power -> moving to VecCore from VC.

Dask: distributed computing for scientific Python

Q. Running on Kuberntes with GCE - can you run it locally? A. Yes, it will run on local resources without any trouble.

Q. Profiler was very nice could it be be ported? A. It can display things from the python profiler from the workers. However, there are probably other more generic solutions.

Q. Scaling? A. Runs on 1000 core clusters, not on 10k, 100k. There is a central scheduler to simplify the programming model, so it won't match the performance of MPIs on massive resources.

Q. Can you share memory, e.g., copy on write? A. People have prototyped it, but it's not part of the standard.

Parallelized tracking algorithms

Q. Does the performance increase take into account the clock speed slow-down when using AXV on Skylake? A. Yes.

Performance and Cost Modeling (Parallel - both sessions)

Overview of the Working Group, Mandate, Activities and Status (Pepe)

Need of metrics in some details
To be translated into cost of the ressources.
How far the impact in the all chain.
    Storage, memory, network
We need a performance assessment
We need to map to cost (global)
Idea: to build a model.

Trying to address:
    one change -> which resource is impacted
    The evolution of the model
Optimise the global cost of the infrastructure.
Many unknown (in kind).

Decision to create a "perf. Systems and cost model" WG (kickoff meeting novembre 9tyh 2017) - 35 participants - developers, sites, experiments
Focus is HL-LHC, WLCG WG, close collaboration with HEPIX, HSF
Long terme activity, continuing process to improve
    Focus: 2019 including most "important" workloads

WG sub-tasks:
    #1 Which important workloads (resource consumers)
    #2 Package these workloads (together with HEPIX) to run them easily
    #3 definition of properties -> input for purchases, developers to improve payload
    #4 draft a cost evaluation: cost to expand the site capacities (workloads -> impact on sites)

#5 list of the relevant performance analysis tools, in the long term build a knowledge base

#6 test beds to run workloads

#7 experiment spreadsheet to calculate the resource needs - create a model that

Sessions : this morning + this afternoon
Volunteer for notes - morning session : Andreas Sc., Catherine, Renaud

Conclusions:

Q- tests beds at sites, test beds on its own or in the regular site
R (Pepe) - idea to have isolated env. To measure applications metrics ; then to compare inside the standard env.

Most important workloads and how to run them (Johannes)

How we are analysing our data -> publication neds several steps
Workflows: simulation stream and data workflows ;
Cost: CPU , memory, disk and network/disk I/O
Categorisation : CPU intensive, CPU+IO, IO intensive

Large fraction of resources go into simulation
Reconstruction Run 4, reco time will take much longer (pileup)

Scale with pile up (slide 5):

Event simu and generation :

Digitisation : input scales linearly, output size, les linearly

…

instructions how to run payloads (python wrapper with experiments specific): see google docs ; all is working fine even with a VO proxy
Measurements of payloads -> goes directly into cost evaluation

Q (Brian UK)- suggest to low at network example of AAA
R- difficult to capture network at the site level, difficult to collect without interfering

Q- data accessible without proxies ?
R- this work once the data are available, f. e. example Johannes provided the ATLAS input by passing the VO protection
Q- shall we put some data in between us in some common space
R- yes if we re given the space

Q- you obtain a number and what ? where are the knobs ? how to tune it ?
R- the knobs is why we are here, to check for example if we create disk contention, network contention ; from the SW side, this is not in our control ; for ATLAS we are running no matter where (except T1 have tapes) ; in that sense

Q (CMS) - in CMS: we did a full chian workflow then break it again to adapt it to sites ; ATLAS intend to do the same ?
R - we can rpovide in 2-4 steps
Q- you did not mention tapes ?
R- we did not looked at it yet ; we have to look into it more ; T1 has to participe and look into it

Q (Jeff) - manpower ?
R (Markus) - this will come later, here reference worloads
R (Helge) reply to Vicenzo : here this is to understand the various cost and then we will optilise ; now understanding
Q (Brian UK) : separation between tape and disk, prestaging ; we have to take into account latency and the impact on jobs efficiency
R (Johannes) - we have to evaluate it
Q (Brian UK) : getting the overall view at the site, in case we are in troubles it would be good to have a view of which payloads are running -> putting a flag on the jobs
R- this could be done, does it fit in this WG, not sure

Metrics and Measurements (Markus for Gareth)

Current approach to define a metrics andlater maps to manpower and machines
We need to characterise the payloads, we can measure a lot if things
We have to do it to guide design and purchase, in term of fundamental capabilities
Manpower is a different stories, no /proc

Impact of changes : a drop in job efficiency at Glasgow -> the job mix has changed
    The system was not designed for this
    If we have know the characteristics of the paylaods, this could have been avoided
WLCG payloads are complicated -> we have to simplificatio, we have to fnd the right cursor

Sample metrics
    CPU
    Memory
    Data consumed/produced

Q: any plan for central monitoring ?
R: avoid to have a large scale plan at the beginning, start working to have the main functionalities for a limited set of workloads, then next step plan if we need some global thing.

Cost evaluation process (Catherine)

Main goal of this sub-working group is to translate resource requirements (experiment payloads) into infrastructure costs, and estimate potential major changes.

We deal with the last step, where we calculate the costs of hardware etc. from payload requirements.
We started with an simple example workload and used it to understand the related costs based on our current infra.
Picked up a payload, a few sites volunteered to establish the costs. The purpose is to understand what is relevant, identify what is missing, find a common approach for an estimate.
Specified data in and out, total block read/writes, processing needs, memory needs.
First, checked that our current infra can handle the payload.
4 sites did the exercise for storage, different costs for storage by a factor 2.
for compute, similar situation.
Not enough to buy boxes, also considered HW lifetime, racks, power outlets, network hardware.
Estimated electricity costs and PUE, buildings, manpower (which is very difficult).
Storage configuration can change the cost, e.g. due to overhead.
Some sites gave integration costs. E.g. they can be 18% of the node cost.
Only one site pays for electricity! Still it is good to estimate the price.
Disk power depends a lot on density. CPU power consumption is more stable with time. To power up disk and CPU, one should add (in this example) up to 30% of the integrated HW price.
We tried to see the gain from decreasing the memory amount (-18% in cost).
We cannot take only one workflow, we have to consider the mixture.
For network, we just checked that there is no bottleneck.
Still many missing things, but we could share prices and costs for the first time with others, it was very useful.
Mentioned also effect of CPU efficiency, licenses, tapes, manpower…
Cost of tape (all integrated) not easy to estimate and might not be so cheap.
Now we are four sites, more are welcome to join and share.
The goal is not to compare sites with each other but learn how and where to gain and how the changes in the computing model for the HL-LHC may impact the cost. Need to collaborate with experiments and resource providers.

Vincenzo: not real money, sometimes some funding came from EU projects, sometimes due to many other reasons. Even if this is correct, still we cannot say that A is cheaper than B. Most of these things are funded as a big block.

Q: PUE ?
R: Power Usage Effectiveness = (IT+COOLING)/IT power consumptions

Photo!

Tools and Techniques for Performance/Cost Measurements (Andrea V.)

Three class of tools. Will talk about open source tools, commercial tools, community tools.

Important to the HEP community to compile a list of tools. Please let us know in case of omissions.
Valgrind, igprof, proc,, prmon, gperftools, vtune, etc.
We have to do system/performance/development analysis.
Not only diff workflows exist but also diff phases in a workflow. Things like valgrind and vtune can also profile at specific times using hooks. For instance in Gaudi you have hooks to make profiling easier.
Perf: main interface to hw counters.
Gperftools: tcmalloc, memory profilers.
GNU gprof.
Valgrind: more a framework than a tool (used by all experiments). It contains many tools (e.g. to discover memory leaks, for memory profiling, thread profiling, …).
Iostat: statistics about I/O.
Vmstat: to collect memory usage statistics.
/proc: lots of real time information.
Intel compiler optimization report: insight into compiler optimizations.
VTune very useful for CPU and GPU analysis.
Intel Advisor: more on vectorization and cache efficiency, roofline analysis.
Prmon: monitor resource consumption of a process.
IgProf: born in CMS, used also by others, does sampling and is much easier to use than valgrind.
MALT and Numaprof: memory allocation and profiler.
FOM tools: similar to MALT, memory allocation patterns.
Trident. CPU analysis, using hw counters.
Summary table with information about the tools (overhead, concerned areas, need recompile).
Lots of powerful tools, each one difficult to master. Sometimes useful to build wrappers to make them more user-friendly. We must identify the most important ones and train people to use them.
Learn how to correlate information from different tools and application-level information.
This is a cross-domain area (application level and system level information), we need to correlate them. It is important for the performances system WG.
Let Andrea knows of any other tools.

Vincenzo: any reasonable malloc has a debug lib providing massive information.
Andrea: is igprof still developed in CMS?
Vincenzo: not in CMS.
Graeme: there are other report on the same subject in other parallel sessions; there is interest in doing this in the sw community and collect docs on how to use these tools and improve the code. Will try to set up an HSF group on profiling, though there is slightly different orientations there is a strong link with your work and we should make sure we collaborate.

Evolution of the critical workloads in Run3 and HL-LHC (David L.)

How will our wfs will change?

How we can fake these changes?
What will be the impact of these changes on the global infrastructure?
…
Detector upgrades defined. Prototype code. How to extrapolate?
Themes:

- Evolution to many cores: how metrics will evolve?
- To small data formats. Tradeoff studies?
- To using heterogeneous resources. Example application models?
- New algorithms? Deep learning? Training facilities?
- New analysis techniques? Ideas for modeling an Analysis Facility?

Need to understand what to incorporate in our WG.

Vincenzo: benchmarks, micro kernels. If we look at tgw workflows, they are full fledged apps. Workload is an assembly of micro kernels. Anybody tried to run HS06 as a grid ?
Domenico: we did run HS06 in the grid. We do it in a worst case scenario (full node). Then HS06 is no more representative of what we get.
Jan: thinking of big workflows as chain if micro-kernels would help a lot this WG. If we can come with a good model of what we do today (each micro kernel), it would be easier to predict the impact of changes.
Oli: what's missing is to look into network. We need to understand the capacity we need and to turn back to network providers. We tend not to talk about non-event data (calibration constants, …), it's not negligible.
David: non-event data in CMS should not change in size.
Pepe: how our apps use the network may change in the future.
Johannes: non-event data in ATLAS should not change either.
Jan: for network we can look at the OS level or at the app level.
Markus: we are already doing both.
Jan: effect of caches can be very significant. Introducing a layer of caching may help reducing IO.
Markus: until we have applications a the future, we could assemble micro kernels and we should try to understand how to build pseudo-HL-LHC applications.

Cost evaluation process, example calculation (Renaud)

- Who pays what?
- Make statistics

What is a cost? Hardware, license, server/rack/switch, memory
Power cost: energy, cooling.
Infrastructure: room, routers…
We use HS06 and TB. Skipping tapes for the moment.
1st approach: estimate marginal cost. Is it enough to prepare for Run4? Not a marginal cost anymore.
Created a spreadsheet that sites can fill.
From a site perspective, it would be good to have an yearly cost. Price evolution seems to be slowing down. Need money to maintain capacity and more money to add capacity.
We can use formulas to modelise costs in an analytical way.

Frank: important to accept that costs are different at different places, by factors of ~2. In my site we are not charged, univ provides for everything to CMS. Same for power and networking. No sense whatsoever to calculate a cost. I'd tell you what I get for free. This should be acceptable. More important is to know which VO is charged.
Pepe: we can still estimate the cost.
Frank: no reason to do it.
Ian F.: what's the goal? I might want to know where it's better to put new resources.
Vincenzo: at high level one may want to know how much money has to be spent, no matter where. There is an established way of account for these contribs for detector construction and operations. This is the goal.
Markus: I disagree. The motivation was that if we see that future workloads have certain requirements, sites need to evolve and need to know what kind of workloads will run and match what they buy with the requirements. The workload description must be translated into needed capacities to size the DC. It helps the sites in buying what they really need. It is irrelevant to compare sites.
Jeff: UPS, alarm system, … For us is 500 euros/rack/month. In the spreadsheet there is no rack, no PDU ect
Catherine: the price in the spreadsheet already includes everything, we have another spreadsheet for this.
Concezio: at CNAF we have also non LHC experiments, for economy of scale. How do we factor this in?
Renaud: ideally there is just a fraction to apply for LHC computing,
Concezio: e.g. at INFN we have some overhead for the disk cost.
Jan: more than sites, we should :

- try to understand what are our chance that it fits for the HL-LHC.
- clarify the confidentiality of the data,
- also useful to optimise cost of manpower (it will be a cost driver).

Markus: no need to expose to the outside world your costs, you can advertise the amount of work you can do within your budget ; sites can compare numbers, but only on a volunteer basis.
Ian B.: we should not expose real costs for WLCG at different places. It is risky with founding agencies. The original motivation is where the biggest costs are and understand where we need to invest our effort to save money. No need to go into too much details. We wish a simple model.
Frank: what is important is the budget needs for the LH-LHC and the tradeoffs to meet the budget.
Ian B.: the answer is: the Funding Agencies want to maintain a flat budget; the cost of LHC computing will be the same as of today.
Frank: for me the central issue is the tradeoff. If I don't improve this, what am I losing on physics?
Markus: you can fit in the budget in different ways. Need to know which throughput you get in each case. If you have no model, you are blind to how many things you can produce.
Vincenzo: the real cost of a T2 is completely different than a similar site built in-house.
Markus: if you get a share of a local computing center, at least you can know the value of it.
Vincenzo: we should not compare different sites.

Pepe: we could still see e.g. that ratios of different cost components are similar.
Dirk: these ratios of CPU/disk/network can help sites.
Markus: if I decide that the experiment's organisation of data is such, I need to know how much it would cost to sites, e.g. how many events they can handle. If I do, I revise the model and I can find the best model and optimise the costs. This is an iterative process. I want a framework and a model for this process. This is joint effort, it is not comparing sites.
Torre: perfectly said!

Common resource calculation model: Brainstorming/Hackathon (Andrea Sa.)

Try to create a common framework for modelling the experiments requests. Started with what is used in US-CMS (the so called "megatable"), refactoring and generalizing it. Get required data and MC event sets as json files, then look up resource needs per event. For CPU presently only CPU time per event, it would be nice to include other things like memory or I/O or software parallelizability. Estimate total CPU requirements and total storage requirements for the input data sets.

Is this useful? Which are the minimal requirements ?How to include network? How to include memory, (remote) IO, parallelizability, ...?

Frank: this was simple by design, chose not to include network, but it would be nice to add it.

Frank: one thing wrong this year (slide 16, analysis extrapolation from past to future), the analysis is a factor of reco, and the extrapolation is not the same considering pileup, it would be nice if you can fix it.

Concezio: this exercise is reflected in what we request from the FAs. It would be nice to attach uncertainties to be presented to FA. Would like more granularity e.g. amount of MC events, full simulation and fast simulation, skimming.
Andrea Sa.: the idea would then to make it more flexible without going wilde completely.
Concezio: yes, this is a risk.

Frank: wrote this because the DOE requested something like this, we wanted to derive scenarios. What is done here: to use it not to derive budget, but to do the inverse, i.e. study how can one stay within budget.

Andrea Sa: difficult for the network to disentangle different experiments and to understand what is going on. A lot of remote access to GRIF.

Frank: we have met with USNET to understand what is the transatlantic traffic. One we'll understand we may make a model.

Pepe: it is important that the network is considered since some sites pay for the network.

Edoardo: two types of network, WAN and LAN. Blocking factors are also important. This parameter can be important, especially in comparing own data centre to a cloud data centre.

Brian: what I would be interested in is how much IO WAN, how much CPU usage to have an idea of the network per CPU.

Pepe/Andrea Sa: difficult to disentangle.

AleDG: the situation is complex. Markus tool prmon starts to provide reasonable information to tackle this. As of today, we don't know what's happening really and we want to understand. also disentangle different protocols.
Pepe/Andrea Sa: xrootdt transfer monitoring does not work well.

Next steps (Markus)

First impressions

- Interesting option to collect information via Condor batch system
- Networks should be more in the focus (the network appeared everywhere)

Very often heard the "what-if" scenario, if we do caching, distant access… we have to start on providing a model to answer.

Hot topic was clearly mapping of resource profiles to local costs. Some confusion

- What is this for ? we have to write it down; importance of complete cost
- How to account for free resources (comparings sites makes no sense)

Common resource needs model:

- more tuning is needed,
- also include the "what-if" scenario
- We have to add network

Still vague: how to describe the HL-LHC workloads ; some ideas emerged like collecting micro-kernels and trying to build different payloads ; same for storage and data movements model

- We need more thinking, a new task should start

Jeff: manpower should be added in the cost model
Jeff: about what-if scenarios, we had some issues recently, mainly because of new gravitational wave community and new ATLAS usage patterns, and no one knew what was happening. We should keep these peaks/expansions of resource usage in the model and the operational cost is huge.
Markus: correct but very complex to parametrise. This cost comes from the complexity of the system and lack of monitoring; no formulae for that, at least we can take into account a complexity penalty.

JA/Markus: We can for instance count the number of systems and attach a probability of failure to each.
Jan Iven: would rather try to use the monitoring data we already have and get info from it.

Vincenzo: for us, developers and experiment managers, what counts is “what we can do with this infrastructure ?”. If we can not afford AOD, we don’t write AOD. A risk assessment should be included.
Markus: like high risk, low impact, for this workload this is crucial or not.

Jeff: large duplication in data sets was mentioned this morning, should include this.

Visualisation (Parallel)

Visualization CWP paper status

- We need help to refine the content for final version of our CWP paper
- Please, fork the repo and push your comments: https://github.com/HSF/Visualization

Project Discussion: Intro

Hybrid Rendering System for Particle Collision Experimental Data Visualization

Ciril Bohak (Lubjana) presented their latest developments in visualization for Med apps, based on client-server visualization and hybrid rendering.
It also implements collaborative tools to share views among multiple users.
Also, as part of a new joint project with ATLAS and CMS, he presented a first customization of the system, to display HEP hits, geometry, tracks.

New development in ROOT concerning web-based graphics, GUI and geometry display

Sergey Linev (GSI) presented the latest developments in ROOT graphics. With the CMS EVE team, they are working on client-server visualization using JSROOT

ATLAS Open Data apps: bringing the interactivity of the website to the desktop

Arturo Sanchez Pineda (ATLAS) presented the recent developments on webapps with the Electron framework, to be used in Outreach / Educational events using the ATLAS Open Data

- Arturo mentioned GitBook as tool. Ed Moyse (ATLAS) asked if GitBook (proprietary) was needed or if other solutions could be used. Arturo answered yes, other solutions can be used to replace GitBook.

VR-based event visualization in Belle II

Leo Piilonen (Belle II) presented the Belle II VR application.

- Riccardo M. Bianchi (ATLAS) asked about the geometry exporters they developed for the project. Leo answered they developed two exporters for the Geometry: Geant4-->VRML2 and Geant4-->FBX.
- Ed Moyse (ATLAS) asked if those exporters could be used by the other experiments. Leo answered they are now part of the Belle II framework but they could be moved out of it and, in principle, be used by other HEP experiments. He will ask more details about that.

Open discussion: Can we build common Viz Tools?

- Unfortunately we had to close the session before its end, because of the Workshop group photo.
- We will continue the discussion offline, in the WG mailing list.

Software Development (Parallel)

Project Discussions: Packaging

- Rebooted effort based previous efforts by Liz and Benedikt
- Not only compiling but also deployment (cvmfs, containers) and development environment
- Important to easily integrate new, standalone tools (hopefully coming out of HSF)
- Use cases document v1.0. to find largest set of common needs
- Created test stack of basic HEP packages for first experiment with candidates.
- Need more challenging stress tests afterwards.

Q: Two main problems: dependency trees. Best candidates at the time, easy build, homebrew not in the list

A: Homebrew failed one of the requirements (original Technical Note).

- One of the main values of Spack is the dep tree (6 FTE)
- Not particular different from what other tools use.
- All tools in the current list should manage dependencies correctly
- Across different architectures/versions? Presumably yes

Q: Incremental build? Why not different packages and keep tool simple

A: Adding extra layers pushes complexity to the groups

- “Modules” within a common environment
- People in ALICE like the fact that they only need to build part of reconstruction package which will be installed automatically
- Same functionality in LHCb, build system of the experiment, not packaging tool?

- Support “real incremental” builds of packages?
- **The actual build system for every package is responsibility of the package.**

Q: what if no tools fulfills all requirements
A: development opportunity to help with the best tool
- Very successful in case of spack

C: More than just packaging system: Nix has different philosophy of building everything
C: Try to use as much as possible standard tools without additions
- Two level discussion: cmake, make, etc on one side; orchestrating dependencies and building all of the packages on the other side

Q: How to deploy, for example in SWAN when compiling on a mac?
- Compile in swan?

Q: 90% of CMS grid jobs containerized … why not do that?
A: On the grid this doesn’t work at the moment … for some experiments?
- Push to containers on grid sites to solve deployment?
- Tools developed to run containers on sites not prepared for containers: Send job together with application. Everything in user space
- Nix does similar thing as containers

Q: Use cases: maybe some tools might have migration costs or additional benefits which might need to be taken into account

Project Discussions: Profilers

- For software to improve we need to make running profilers trivial: If users need half a day to get started they won’t have time
- Continuously profile the code over time to spot regressions early.
- Comparison between tools complicated, can be done
- LHCb has a crude python framework to use callgrind, perf, igprof, common project, starting point

C: Difficult to compare profiles in different run, maybe a tool to “diff” profiles.
C: Training is helpful to explain.
- Mentioned in starter kit
- Lots of little bit of pieces of documentation around (LHCb)
- Collect and combine all the documentation, push upstream where possible

C: Make it easy to share profiles? Web interface like a gist/diff

Projects:
1. Create consistent common documentation
2. Abstract data model to compare different tools -> Go profiling format?
3. Web interface to show differences (maybe based on the abstract model)
4. Regression test suite to spot degradations early on

Giulio will create an HSF group and send information around

Project Discussions: Static Analyzers

- Most tools based on llvm/clang, clang-tidy as most suitable candidate

- Common tools, common documentation?

Q: I'm surprised you didn't mention coverity. Very complicated analysis features but not extensible. It does have a database.
  - Coverity requires quite a lof of infrastructure and is very complicated for the developer.

Q: Standard C++ guidelines we all adhere to
  - Can be integrated into clang tidy

Q: Why not standalone clang-tidy plugins?
  - LHCb has a hacked version which does not require modifying clang tidy

Projects
1. Common documentation
2. Convince clang-tidy to allow standalone, custom plugins?
3. Common set of plugins?

Combine with the previous list.

Project Incubation

- How to get started on projects?
- How do we stop when it's not worth it?

Q: Recognition is hot topic in LHCb. Find a mechanism to provide quotable reference?
  - Peer review and then publish with zenodo. Provide tests, provide release, editorial review.
  - Can HSF give extra credits?
  - Have HSF github repository?
  - Problem is to make work recognized for universities + funding agencies

Q: Technical observation about cvmfs: now we have multiple-user repos where multiple users can write in their own directories.

Q: Deployment: do we need to work at lab infrastructures?
  - Heroku not free
  - Resourcer should only need to develop his code

Becoming "Mainstream"

- Invite outside speaker, form connections
- We might be invited back

Q: Lot of projects don't encourage contributing to OSS because then it's not part of the project.
  - Goes back to recognition

- OSS contribution career advantage for industry, why not science

Out of time so no questions possible.

Simulation (Parallel - both sessions)

LHCb - Preparation for Run3

-
- Simulation will be dominant in Run3 -> need for speeding-up
- Developing an integrated sim. Framework, allowing to mix components/simulation flavors
- Fast parametric simulation: goal 100-1000x faster than full, deploy in steps to allow user analysis to adapt
- Fast Calo sim - aim for 3-10x speedup of calorimeter simulation, looking into GAN options
- Geant4 10.4 validation ongoing. Biggest time consumer. No speed-up observed using VecGeom when moving to this version.
- New developments: parallelize Gaussino,
- Maintain Run2 support while developing new features for Run3

Q. Gabriele: Speed-up not observed with VecGeom due to usage of simple shapes. CMS uses more complex ones. Moving to 10.4 will bring extra improvement

Q. Kevin. Fast sim will be fed back to Delphes?

A. Yes, expectation is that fast parameterisations will go back

Q. Witel: Using G4 hooks for fast sim? Use of GFLASH?
A: We do use G4 hooks
A.No use of GFLASH

ALICE - Preparation for Run3

- Common O2 framework - re-write of the framework with online processing of most offline tasks
- Particularly large events, putting large CPU time and memory pressure
- R&D: sub-event parallelism, profit from monitoring data, profit from new R&D (vectorized initiatives)
- Splitting event for several SimWorkers -> speeds-up and enable a parallel simulation of large events. Works across simulation engines and heterogeneous resources.
- Integration of VecGeom navigation in Geant4 - prototype ready (G4VecGeomNavigator)

Q. Pere: What can we do for fast simulation in common
A: Machine learning techniques, common components

Q. Sub-event parallelism, how much intervention on G4 level. Navigator in G4, same as in GeantV

A. None at all, transparent. The navigator is different than in G4, but dispatch is the same.

ATLAS talk

- Full sim = 39%, will have the same trend in Run3. FCAL biggest consumer, improvements on the way
- Faster or simpler: faster geometry solids, simplified geometry, aggressive cuts, russian roulette for neutrons
- 4% gain from solids, 20% from static builds or 10% from having a single shared lib
- Geant4-MT + AthenaMT

Q: Witek: Had x-sections - numbers for gains due to tabularization
Q: Daniel: Plan of flexibility of chosing different approximations continuing, more options.
A: Using this flexibility, but maybe not for every campaign

Geant4 R&D programme

- Sub-event parallelism: each primary particle or group of them will be seen as a sub-event
- Too many generalized classes e.g for ions leading to expensive 'if' statements -> redesign
- Pushing production cuts to processes -> simplification of tracking classes
- Specializing transportation e.g. for charged/neutral tracks, VecGeom navigation, working better with parallel geometries
- Track-level parallelism - understand costs in this approach
- Moving to Git this year

Q: Paolo: GPGPU prototype non applicable to HEP?
A: No real expectation to apply to general purpose simulation workflows, but just to specific ones.
Q: Andrei: What answers will provide track-level parallelism simulation in Geant4 extra to what GeantV prototype will?
A: Some of the impact of vector of particle workflow can be theoretized
Comment from Pere: we should not just bring up difficulties in using common tools, but be more positive and work towards overcoming them

Electromagnetic and Hadronic Physics requirements for detector simulation in 2020s

- Substantial progress in EM processes and Msc
- Strategy: Aligning with needs of HL-LHC
- Reporting improvements in many models, EM & HAD
- Roadmap including revising code for better precision/performance

Q: Tuning to exp. results - any optimisation framework to do this automatically?
A: Daniel: Ongoing effort to study the behavior of models with respect to parameter changes, triggered by neutrino experiments. Marc: Should expose this requirement to Geant4 technical Forum

Q. Pere: Not re-write EPOS but first validate.
A: First attempt first to connect to Geant4

Towards modularization and vectorization of Geant4 physics: a pilot study

- Modularize G4 Bertini model: pilot project. Motivation: one of the predilect hadronic models for up to 12 GeV
- Removed dependencies to Geant4, turning off some options of activating other Geant4 models, no MT. Can run as backend to G4 simulation or standalone.
- Profiling: pre-equilibrium takes 90%. Optimisations in several places lead to 13-25% perf improvement
- Ongoing work for supporting vector particle mode as well

Q: Marc: Ignore MT mode: no planning for thread safety?
A: No need to use specific MT tool by itself, but code is thread-safe

Status and plans for the vectorised particle transport “demonstrator"

- Status of scheduler - basket flow to all components, user interfaces adapted to MT flow, close to Geant4 style, but with more complex user actions due to mixing of tracks from different events.
- Concurrency adapted to work with external frameworks, magnetic field propagation vectorized
- EM physics validated against Geant4, a list of examples available
- Ongoing work on vectorizing physics with promising results

CmsToyGV: Integration tests of GeantV in CMS software framework

- Toy-mt-framework included in GeantV + example, used as base for extension
- CmsToyGv example: cms2018, uniform 4T field, EM physics. Feedback between GeantV and experiments led to important adaptations in GeantV
- GeantV as external package in full CMSSW - improvements in GeantV using full CMSSW features ongoing. More adaptations needed to allow full tests.
- Full program of tests foreseen + a plan for possible future integration/adoption

Q: Pere: Did you run with several events in flight? Overheads, scalability?
A: Yes, any number of threads is possible, no significant measurements of performance done yet, but to be done. Some speed-up of 7x on a machine with 8 cores observed however in the ToyCMS example

VecGeom in CMS, Muon g-2, Mu2e

- Tests of Geant4 + VecGeom as geometry backend were started early 2017 in CMS. Several issues/bugs found in the process of testing, leading to a production-quality version.
- Physics results validated and speed-up of 7-13% observed for CMS.
- Plans to migrate for Muon g-2 and Mu2e to Geant4 10.4 + VecGeom very soon

Variance reduction techniques in HEP for Geant4

- Introduction to biasing techniques for rare events, biasing examples in HEP
- Could we apply biasing techniques to production of background ? Given the filtering done by the trigger+analysis, remaining bgk events are “rare events”. Focusing the CPU onto -eg- the phase space region at generator level would enhance the statistics of the final bgk sample and would reduce the portion of bgk events just eliminated by the analysis. Advise of physics analysts much desired. Biasing detector simulations in HEP?
- Enumeration of biasing techniques in Geant4

Pere - experiments are doing biasing since they recycle minimum bias events, hence doing a kind of splitting.
Kevin: analogy with applying cuts in the phase space.
A: cannot just cut since information gets lost, using biasing moderates the less interesting part of the spectrum without loss.

CaloGAN: a novel physics simulation with Neural Networks trained from Geant4 full-sims

- === hard to get anything: bad audio connection ===

Q: Pere slide 15: distribution non-isotropic
A: ?
Kevin: slide 17: bias in some places - can we understand why?
A:
Q: Witek: what events you used for training? How do you go from single angle scenario to general realistic approach?
A: Train on millions of particles with different parameters - introduce these parameters in the inputs

Use of HPC resources with Geant4 simulations

- Describing path to using HPC
- Describing options and results for the different parallelism options
- Scalability to million threads on the reach with G4MT+MPI, but I/O to be kept under control

Q: Andrei - pthread and TBB may not be at the same level, one is a threading technique and other a scheduling system. Is there any plan to support task-based workflows.
A: already there, trying to simplify the approach and use
Pere: What needs to be run in HPS is a real application.
A: Using full CMS geometry and full physics, with no digitization that. ROOT not able to handle the scale of the problem -> use smaller number of threads, but the result can be extrapolated. Pere: How experiments benefit from that A: CMSSW and Athena not running in the HPC environment yet.
Daniel: projection for future HPC?

Andrea: not being able to use accelerator, HPC will have them. Worry: cuts in Grid funding in US, while new technology not friendly for our code.

First steps on simulation readout vectorization: LArSoft

- Profiled LARSoft to decide where to use VecCore, promising preliminary results

Q: Sandro: plan for full vectorization of TPC digitization?
A: not yet, now driven by profiling. Next: signal processing in reco side before.
Q: Kevin: consideration for future design, what recipe to give to users?
A: This is for the future

Security (Parallel)

Trust & Policies

Q: What about GPDR and LHC experiments rather than WLCG?

A: By WLCG I mean all, including experiments. Their services must be compliant etc.

Q: What happens if ancient experiment service hits something?

A: Have to balance importance of the service vs the risk. Generally risks very small. Authorities after big bodies doing things (eg recent press coverage.)

C: Need to be able to justify why you're doing it. Probably even true for an ancient service.

Q: Is this the end of delegation? Since we link to credentials elsewhere?

A: Approach should be to see why we are doing things and justify that, and inform people what is happening. Where their identity might end up. Sites running jobs, Amazon cloud etc.

Q: Are experiment security contacts involved? Working on this? A team?

A: No magic team of people. Had a big push in the MB to make everyone aware last year.

Authentication and Authorisation

Q: Capabilities vs Roles. Globus had the CAS. Didn't scale, hence VOMS.

C: Agree. Important to give ability to those deploying to choose the model. Some will need more central, some more fine grained and local.

C: "Who is allowed to use a MC request?" Roles. May need capabilities say if staging: can this WN do staging.

C: History of grid is sites ceding "namespaces" to experiments. Eg Pilot is a role seen by

site, but within that “namespace” the experiment decide what payload jobs run.

C: Need to be able trace who did what in the future somehow still.

C: We do need to be careful to weigh up the cost of re-engineering existing services with new credentials rather than just using a token translation service

C: We may have a 20 year transition period.

Operational Security

Risk landscape for the next 5 years (prep Romain)/Incident Response & Traceability: Vincent

Q: Have you had any pushback from operational teams on cooperation?

A: No, not seen that. Good collaboration.

C: We should have an ingestion layer which checks containers submitted by users.

C: What if users then do a yum install in their container?

C: You don’t allow users to install anything after start time. You insist everything is put to an upstream and then check there.

C: That undermines the flexibility users want from running in their own containers.

Technology and Threat Intelligence: David

Q: Threats from grid sites have been in the context of user jobs. Could this work extend to threats from compromised services? Eg centrally deployed experiment edge services.

A: That’s a very good question. One area is how we form the trust groups. Also need to consider new usages. Looking for trust groups that don’t yet exist. Who is it that you need to trust that you don’t yet?

Q: Start this by sending an email to the list? Contacting David.

C: Want to highlight that grid sites are not threat to general site (campus). In reality the host site is the threat to the grid site. Often a firewall between grid and general site security. What do we have that they don’t have, which we could offer? Eg lots of info about what is going on round the world. One way to build trust.

C: With virtualisation and containers, a lot more people are doing what was a sysadmin role. The channels we've created for contacting site admins are sufficient for this and need to be updated.

C: Work in South Africa to use info from probes to work out what warnings to send to site admins about security.

C: Please join the relevant mailing lists and join in!

Discussion

Discussion in line with talks

Frameworks (Parallel)

Overview of Serialization Technologies

Q. any reason not to use ROOT for file apps?
ROOT fine for that, other use cases (message passing) not a strength
ROOT is not language agnostic

Programming languages for frameworks: is C++ still the best choice?

Any change has to be via some period of interoperability - old code will live a long time
C++ offers a lot of backward compatibility (good!) but bad code and wrong concepts will also still be supported (bad!)
Does it behoove us to use a better language for core concurrency rather than something that's inherently dangerous? Large community of non-professional coders - discipline is had. Enforce a subset of C++ so that only the healthy subset of the language is allowed?

Discussion on Frameworks Section of CWP

Work on terminology - largely supported as an idea. CHEP04 paper that might be interesting to look back too. Also useful for data preservation.
Review of the concurrency model? CLAS and ALFA solve a lot of problems that are not solved in Gaudi and CMSSW. This seems to be good to look at again. GPUs and accelerators are going to be essential for us in the future. Frameworks deal with a lot of data now - batching data for offload is an essential part of the model for ALICE.
Actors and message passing are a lot more friendly to multiple language implementations as well.
Data model will require substantial reworking for any message passing model to avoid a lot of overheads. See Jim's talk about data formats more suitable for intermediate passed/data types. Need zero copy. Eventually have to go to a final archival data model (ROOT, probably). R&D?

Have a fwk workshop in 6-9 months? (Probably US?) November? A resync at CHEP (Sunday PM) could be done - BOF.
Can we have more regular meetings/updates a la concurrency forum? There is some effort in the context of GeantV (Chris Jones, Andrei Gheata) that will happen. Anything else funded? Not that we know about - but perhaps we bear this in mind to have more checkpoints. **Volunteer for a convenor in this area?**
DSLs - not clear that this would be that useful for us or that we have effort to do that right now.

Common Data Management and Data Lakes: Technical Session (Parallel)

Distributed Storage and Data Lakes: ideas from EOS experts

Current dist storage: several sites with disk, each running a deamon connected to centralized namespace (WebDav, XRootD, S3 or filesystem), no API extension to external storage. Common API needs to be defined as well as data distribution (tuning parameters).

Evolution: attaching external storage with local namespace instead of as object storage -> extended connector API (Hybrid storage, multipath gateways…)
Dynamic caches: read/write cache transparently connected to lake. XCache requires some add-ons.
Big pict: Interface for High Level DM systems and data placement

Q: Access over WAN
A: Ask the closest piece, than synchronizes

Q: CEPH is slow for metadata, is it the same for EOS?
A: It’s only inefficient for small, multiple updates

C: (Mario) Very much in favour of the common quality interface
C: QOS interface might be an overkill. A list of file needs to be accessible for high frequency queries
C: QOS interface decides what should be done at what cost, like a staging command
Q: (Brian) What happens when experiments decide for a huge change, is there prioritization foreseen?
C: (Oliver) FTS can be extended for quality of service
Q: (Frank) Three structure of accesses was removed? Will it be re-established? How will the closest data be found?
A: With right-once read-many local access and lake if not accessible locally is possible

Distributed Storage and Data Lakes: ideas from dCache experts

Current status: All protocols, tape support, replicas are fully configurable, data locality preference
No central way to update system, all data servers need to be dCache, no operation

Future: Non dCache components as data servers, internal locality, cache sites (read-through, write-back/through). Protocols: S3, swift plus locally installed and CEPH pool support

Q: (Mario) Can we get QOS interface
A: It's there for 2 years
Q: (Brian) NDGF and ATLAS setup are run by ppl paid from the same source?
A: All the pools are run by institutes, downtimes need to be coordinated, especially pool downtimes.
Q: How would operational model deal with scale-up?
A: 2-4 times should be feasible. Synchronizes downtimes for all pools is difficult
Q: There is preferred local access, what is it is not available?
A: It can be decided if data is copied or only retrieved. To be checked if cache is possible. One pool does copy, the other doesn't as it has low latency cache.
Q: What would happen if r-cache was switched off?
A: There is an impact of remote reads. Could be mitigated by copy to collocated pools.
C (Shawn) sites of same sizes, so files can be temporarily copied to other site. There 8% space left as cache

Q: In HL era, EXA-storage, how would the storage work
A: (dCache) I'd go for federated site with shared namespace accessing a part of the tree close to you.
(EOS) It's hard to project, needs to be measured. High rate is a different issue than data size
C: (Ian) A project on how distributed project can work is proposed to understand how it can be done. There was a proposal already in 2014 but in was not founded. Daisy dCache was a part of that proposal. It's R&D to understand how it can work.

Distributed Storage and Data Lakes: The XDC Project

EU founded project to develop scalable federated storage by adding functionality to existing projects.
Q (Brian): There are multiple project here, cross-project discussion would be useful. CMS as a community would like to be there for requirement gathering.
A: A change in the way project is organised is needed.

The ESCAPE project

Cluster of big scientific projects (HEP and astrophysics) to shape EU's open science cloud.
Goal: prototype infrastructure for Exa data needs of those communities. Focus on FAIR DM

(Data Infra of OS is lead by CERN with a number of data centers involved/financed). Commercial cloud access and network work will be included. This is an integration project, not SW development.
Q: Will Rucio be included? Will there be QOS interface?
A: Yes, possibly

Using Capability-based authorization for HTTP-based Third Party Copies

HTTP headers -> how the public key is found and recognised (iss, issuer) -> Token verification -> Token authorization
Implemented already for Xrootd
Token comes from FTS x509 proxy or from a fallback strategy (DCache macaroon or ir Grid Based)

Q: Transfer layer. Rucio is issues for Atlas and submits request at the same time. Is it possible to submit directly to FTS without a trip back,
A: For the moment FTS doesn't support it, queuing is also an issue (fresh tokens will be needed). The trip back is important.
C: dCache will issue macaroons to Rucio and than Rucio will contact FTS and FTS will issue another one. The whole stack is still needed.
B: Macaroons have different properties, they will be good to combine
Q: Is the VO token used as membership token and target token as real authorization
A: Right boundaries need to be found. For example for coarse grained to fine-grained. Two approaches need to be combined.
Q: Standardized token interface will be needed at end-points
A: It's not there for the moment

XRoot and HTTP Third Party Copy

No standard way between protocols to do that but they are widely supported.
A free replacement needs to be found when globus/GridFTP support from WLCG will finish in 2021 (or earlier).
XRootD proxy server is functional and was tested by Atlas. Proxies can be deployed in pools. More evaluation is needed.

Q: GridFTP is a standard, globus just drops implementation. Should we take a non-standard?
A: Atlas needs something that is not costing money. HTTP is also standard.
C: Storm is missing from tables. Can proxies by deployed on existing hosts
A: Yes
C: Site wouldn't like to have additional traffic from proxies
C: (Brian) With HTTP approach transfer protocol can be different. Possibilities should be explored
Q: What are the plan to publish
A: For HTTP HTTP copy is used by default in Xroot server.
C: (Wei) There will be management tools necessary for the new tools:
A: Yes, it needs to be done during evaluation period.

caching technologies: Xcache

Xcache is cache for static files, based on plugin architecture.
Used by Atlas and CMS. In RAL works with CEPH with XRootD plugin, in BNL as squid replacement (supports HTTP and xroot). Can find closest data.
Future dev: HTTP support between cache and data source, Cache eviction for general purpose cache. Read-write support is considered
How can Xcache be used for HPC effectively is to be seen.
Q: (Maarten) How big are the caches at site?
A: In SLAC 22TB on one machine, UCSB is 10 times bigger
Q: It is significantly smaller than storage
A: Yes, but it can scale. It's more for analysis than production.
C: (FRANK) Max amount of spindles, as little space as possible to provide speed.
Q: (Oksana) Is cache a part of storage or an application? Who should manage it?
A: Advantage is that the cache is managing itself.
Q: Proxy mode and cache mode was tested. Can be used as a IPv6 proxy to IPv4 storage
A: It requires further discussion
C: (Frank) For many small communities self cache management is not a option due to management overload.

caching technologies: HTTP caches

Many scenarios for cache performance benefits. Based on Volatile Pool in DPM. Dynafed was used for WebDav endpoint aggregation. Dynafed redirected access in the setup model and provided availability in case cache disappeared. Massive stress tests are still needed.

Q:(Jan) Since cold cache is worse than no cache was the hit-rate investigated for Bell files?
A: Basic functionality is now tested.

Datalakes parallel session (second set of notes)

<u>Andreas - ideas from EOS experts</u>

- Present different scenarios for distributed storage setups with a central namespace
- Emphasize QoS importance for cost saving in datalakes
    - File conversion
    - Media tape(endpoint/site)
- Need for a common API for QoS management
- File placement co-located versus scattered with RS layouts
- Deployment based on access tokens, protocol redirection suffices.
- Allow stand alone storage (lake bypass) (=> enable notification API, ie. inotify, IF in xrootd, etc.)
    - Allow AWS workloads through notification mechanism
- Multipath gateways (big storage): usage of proxies, every FTS can be acting as one.

- Evolution of EOS: site draining/balancing, probes (heartbeats) black/whitelisting of sites
- Dynamic caches: attach a cache to the lake, replicas in static (eulake) and dynamic resources (cache-foo, xroot cache)
    - Write-through, delete-through caches, credential tunneling, contains notification mechanism (see changes on the fly)
- Common: notification mechanism, integrate different platforms, QOS interface, mechanism to attach a cache.

Q: Rob gardner: erasure coding over the WAN, comment?

- 300MB/s limitation right now, one client collects.

Metadata service is the bottleneck?

- No, in CEPH is like this, not in EOS if you put big blocks as data is streaming.

Q: Mario Lassnig. COmmon interface should come fast.
Q: Jeff Templon: cache is a tidal pool. SUggestion how to expose the common interface (namespace)
Q: Frank Wurt. Choose opposite strategy than Torre. Fabric decides who goes to tape?? NO. It is the experiment deciding.
Q: Brian. Comment about the common interface for QOS. Oliver reply that FTS is exactly doing this already.
Q: Frank. Will you reintroduce tree structure that got lost - closest replica? Andreas says this is the hybrid model: local and then lake access if copy is not there.

<u>Tigran: ideas from dCache</u>

- Describe dcache deployments around the world
- Describes internal components of dcache
- CAP theorem, we can have all (but network failures)
- Showed data lakes and claim they did the same long ago. Claim that we need to talk to them to know what to do and how to run a system like this.
- Missing in dcache: all data servers must be dcache pools.
- Future:
    - mature NDGF and AGLT2-like deployments
    - Use non metadata dcache servers
- Pure caching:
    - cache warmup concept
    - Pure caching devoted sites
    - Ability to turn a site into cache and back
- Usage of non dcache storage blocks
    - Object stores
    - Use local storage: EOS,DPM, xrootd
    - Preserve namespace
        - Done for CEPH
    - Avoid site harm

Q:
Mario Lassnig: can we get the QoS interface? Tigran say is there since 2 years.

Q:
Torre: NDGF and AGLT2 setup they are all under the same admin domain to work together, is something we need to break?
Matthias: all dcache pools are run by local dcache
Torr: Need to coordinate downtimes.
Matthias: yes, need to.
Torre: do you see model scaling up with more sites?
Matthias: I could see absorbing a factor two.
Torre: Operational models are part of the key…
Tigran: make sure consistency taking or a dist datasereervs.
Pepe: slide 12. Preferred local access? When data not available, auto-replication (Tigran say is configurable) Arc cache setup is in NDGF not in AGLT2,. Pepe: if ARCcahe is switched off what about the perf? Matt: did some tests and there is an impact.

Q:
Ron Trompert: 20B files and 3EB how this can be implemented? Tigran: federated sites, single namespace and multi-sistems, kind of re-direction like model. Andreas: numbers are not important, iops is the number that matters…

Q:
Ian: go back to the R+D storage definition this is what was presented. Nobody said that will be like this or like that, is an R+D to see who we can proceed… we are trying to understand, not imposing anything.

<u>Oliver Keeble: XDC project</u>

- EU funded project just started 3M
- Run large distributed data infrastructures
- Improve already existing IF: indigo, onedata, fts, EOS,etc.
- Is a usecase driven project: ECRIN. WLCG. XFEL, LifeWatch, CTA
- XDC topics to attack
    - QoS
    - Datalakes operations: caches, compute integration
    - Cross-federation data management
- Claim homogenisation of data lakes concepts, need to agree among all of us to refer to the same thing.

Q: Brian: need to talk to experiments.

<u>Ian: ESCAPE proposal</u>

- ESFRI infrastructures clusters of science come together.
- One cluster is particle physics and astronomy: HL-LHC, FAIR and SKA, CTA, KM3….
- Goal is to prototype an infrastructure for EOSC adapted for Exascale data management.
    - Oportunity for funding to work on the prototyping of a datalake.

    - Main players: CERN and DESY (EOS and dCache)
        - Aim to integrate different storage systems since the beginning
    - Coordination with GEANT on networking aspects
    - Auth/Authz
- It is not a development project is to put all parts together: datalakes, XDCs, FTS,...

Q: MArio./Ian, yes Rucio can be part of it

Brian: Using Capability-based authorization for HTTP-based Third Party Copies

- Data transfers with http over fts.
- Ideas for public key verification for data transfers
- Usage of tokens for transfer layers.

Andrew Hanushevsky: XRoot and HTTP Third Party Copy

- Third party copy review
- There is no standard for TPC:x509, transient token, etc.
- Man HEP protocols support tpc: gridftp, http and xrootd
- Globus eneden support for open source, incl gridftp ($$$)
    - Support promised until 2021… but….
- Free replacement should be found for gridftp
- TPC should work for all combinations of storages, protocols, FTS and Rucio
- Xrootd proxy is LHC storage interopreable
- CASTOR and EOS does not supprort WebDAV
- CASTOR, DPM, EOS vanilla xrootd ready
- Dcache requires xrootd proxy
- Xrootd support multi DTN (as a cluster of proxy servers)
- Conclusion:
    - We do can replace gridftp
    - HTTP and xrootd widely deployed
    - Free data management looks doable

Wei Yang: caching technologies: Xcache

- Summarized Xcache activities to optimize access for: CEPH, CVMFS, RUCIO, ESS.
- Detailed setup at RAL for xcache setup with CEPH
- devel :
    - Core devel ended one year ago
    - Xcache can do http between clients
    - Cache eviction
    - Read-only or Read-write
- Packaging and deployment plans presented
- Describe why Xcache for HPCs is not a good idea.
    - Q: Maarten: how big are these caches? Wei: Need to be scaled accordingly Frank: maximise spindles, minimise space (11 nodes with 12 disks each disk 2TB)

    - Q:Oxana: is xcache is part of the storage or the application? Wei: cache is managed itself… is a pure cache.

Silvio Pardi: HTTP caches

- Present SCOReS project - study of caching systems
    - Setup and test a caching system to integrate in experiment computing model.
    - Pilot experiment is Belle-II
- Started cached investigation with DPM
- Describes the concept of a volatile pool
- Described Volatile pool with Dynafed
- Showed the testing phase progress

Thursday

Common Data Management and Data Lakes

Comment: Be wary that interference between components can create an entangled messy system, like we have today!

Frameworks and Infrastructure

Comment: we might need task based and data flow models - it's not one or all.
Q. What are the plans moving forwards? Yes, we have to follow up - maybe a 1/month meeting to keep up momentum. (Open point - *make sure this happens*!)

Programming for Concurrency and Co-Processors

Q. Very good talks - how do we go forward from here? A. Projects are going and they make sense on their own. Common themes are there.

Simulation

Comment: For HPCs it's the experiment embedded simulations that run, so testing these is much more important than toy applications. Response: Also cooperation with facilities. Tests done at Mira were quite realistic.

Analysis Facilities and Use Cases

Key point is how to organise going forward to tackle CWP points.

Software Development

Q. clang-tidy discussion, did we discuss what tests we should run (CMS now run nightly, discussion on options was needed). A. No, didn't have time to go into those details. We should make sure this knowledge is easily available.

Visualization

Impressive collaborative effort!

Workload Management

Comment: Usage of tape - there is a WLCG Archival Storage Group looking at tape (Oliver chairs and organises, next meeting in April). Generically look at use and QoS for of 'cheap' storage, that may not even be tape in the future.

Security

Q. Reach out to broader community (campus) is good. Are there efforts to reach out to the private sector? A. Threat intelligence is a critical area and it helps build trust. A threat intelligence setup at CERN does have many relationships with private sector and government. We want to build these frameworks more widely.
Q. Remote participation - did you get that from campuses and other grid sites? A. Yes, we did. There is also a workshop at the end of June that will strengthen links.

Performance and Cost Modeling

Q. Have you considered a simulation tool to help with overall computing workloads? A. We have been involved with some people who did discrete event simulation. This has not been popular in our community, but we need someone to work in it as it can be useful. However, it needs input parameters that are not available for HL-LHC.

Data Preservation for HEP

Q. CWL (common workflow language) is used commonly outside our community. Would it be something that will grow and could we use it. A. It is growing and has a large community. There are competitors and it does abstract from all the details very nicely. It's probably the best bet and gets a lot of other inputs.

Training and Careers

Q. Careers are very important - we keep losing people from the field and this has a very detrimental effort to software. A. Understand and agree with the point.
Comment: Encourage and advertise people to come in and that there are HEP opportunities outside as well (we cannot keep everyone). We need both.

Comment: 85-90% of people will move on, change that people can get faculty positions where software development is recognised.
Comment: Bigger issue, historically S&C effort has been a small fraction of the total funding. We need to enlarge that fraction and its recognition. Cultural change that we must convince the community of.
Comment: Software is publishable! Make sure your software contributions are counted. That can help people to stay in or close to the community.

Conclusions and way forward