OCP HPC Sub Project

Minutes - 10/01/2024

OCP Local Recording :
https://drive.google.com/file/d/1eudgdFT26YQcz5ODqNBvX-4F0RYFa53E/view?usp=drive_link

Meeting Transcript :
https://drive.google.com/file/d/1rPpkPBBOfw-PQ0q7tHsq-XpSJtLxdStf/view?usp=drive_link

Attendees

Attendee	Affiliation	10/01/2024
Munir Ahmed	Lattice	x
Raul Alvarez	Eisar	x
Kevin Cameron	Individual	x
Allan Cantle	Nallasway	x
Michael Choi	Samsung	x
Robert Ciotti	CTG Federal	x
Zhineng Fan	Amphenol	x
Olivier Franza	Intel Foundry	x
Bobby Lu	Lightelligence	x
Geoffrey Mersham	Unify Point	x
Mehmet Sencan	Atlas Computing	x
Totals		11

Meeting Objectives

Proposed Agenda for this meeting :

1) Introductions to New Attendees.
2) Discuss DOE New Frontiers Proposal
3) AOB

Meeting Discussion Notes

1) **Introductions to New Attendees.**

Olivier Franza - Intel Foundry - Leading a team of system architects in intel Foundry - aurora Supercomputer Chief Architect. Working on next generation racks and confiugations and part of the OCP AI initiative. Out of the discussion with Bijan Nowroozi I found out about the wall of compute. Aspect ratio of the current blades are not ideal. Recommended him to talk to Allan and discuss this thread.

2) **Discuss DOE New Frontiers RFP Response Proposal**

   Link to proposal : https://docs.google.com/document/d/1fsPapI4O4573B6eQM9tnNxYyQymrIZ3dt0RNsYV4flk/edit?usp=sharing

3) **AOB**

AI Generated Meeting Summary - NOTE this is an OK summary but is confused in several areas.

Meeting summary for OCP Server - High Performance Computing (10/01/2024)

Quick recap

The team discussed a high-performance computing (HPC) wall of compute proposal, focusing on a pragmatic approach to keep costs down and improve chances of success. They also discussed the development of an FPGA-based accumulation platform, the progress of three innovation topics, and the potential for collaboration with other companies. The conversation ended with discussions on the implementation of a new system, the use of HPCN, and the potential applications of CXL.

Next steps

• Allan to send Olivier links to the OCPH Wiki and previous presentations to get him up to speed on the project.

• Allan to update the proposal document, potentially moving the full concept description to an appendix.

• Geoffrey to add boilerplate content to his assigned sections of the proposal.

• Geoffrey to send Allan the budget template spreadsheet.

• Allan to add the budget template to the shared folder with other proposal documents.

• Team members to review the pragmatic scope outlined by Allan and provide input within the next 3 weeks before the October 21st deadline.

Summary

Casual Greetings and Discussion on Storm

The meeting began with casual greetings and discussions about the weather in Texas and the aftermath of the Helen Storm on the east coast. Munir shared his surprise at the severity of the storm and its impact, with over 130 people reported dead. The team also welcomed Olivier, who was participating in the meeting for the first time. The conversation ended with Allan preparing to capture the discussion and Olivier looking forward to his first participation.

HPC Wall of Compute Proposal Discussion

Allan led a discussion about a high-performance computing (HPC) wall of compute proposal, emphasizing a pragmatic approach to keep costs down and improve chances of success. He also mentioned starting a document for team contributions. Olivier, the chief architect for the Aura supercomputer and a system architect at Intel, expressed interest in the HPC world and the wall of compute initiative. The team agreed to work on a document offline to contribute specific pieces of work. Allan offered to send Olivier short presentations from the summit for better understanding and discussed the challenges they are facing and their plans to secure funding from the DOE. The due date for the project was confirmed as October 21st.

FPGA-Based Accumulation Platform Development Discussion

Allan discussed the development of an FPGA-based accumulation platform, focusing on building a practical proof of concept within a two-year timeframe. He introduced the HPCM module, which includes an AMD FPGA and additional LPDDR memory for simulating HBM, and proposed leveraging Corono's mini chassis for the HPCM modules. Allan also suggested the HPCM module could serve as a chiplet enablement platform and mentioned working with Cornell to potentially include 36 HPCM modules in the design. He emphasized the need to keep the project simple and within the agreed-upon scope, and suggested pushing certain details into an appendix to avoid confusion.

OAI and OAM Module Progress and Future Plans

Allan and RCIOTTI discussed the progress and future plans for OAI and OAM modules. Allan mentioned that there has been a significant step change in Nvidia's approach, prompting a need for a new concept. RCIOTTI questioned whether the proposed concept was fully baked and if the budget of $2 million was sufficient. Allan acknowledged the need for an elaborate thermal solution, including 2-phase capability, and identified three compelling pieces: power delivery and thermal management, universal interconnect, and contributions over the last three years.

Innovation Topics and Potential Collaboration

Allan discussed the progress of three innovation topics: a system management piece, a domain-specific platform, and an FPGA/ASIC emulation use case. He mentioned the challenges faced in getting traction for these projects and the potential for collaboration with other companies. Allan also expressed his pragmatic approach to the projects, focusing on the FPGA/ASIC emulation as a feasible starting point. Unknown Speaker introduced a European startup, Openchip, which is involved in chiplet AI accelerators and is considering joining the OCP community to support these projects.

Creating a Demonstrator for Compute Approach

Allan wants to create a demonstrator that looks like a wall of compute to showcase the approach they are pursuing, despite its lower power consumption due to conduction cooling instead of water cooling. He emphasizes the importance of making progress and not falling behind the industry, while also acknowledging the challenges of getting different industry silos to collaborate. The discussion touches on the idea of using existing Oai work for FPGA emulation, but Allan rejects it as it would not align with the desired vision. The conversation also covers the collaboration between different groups like Ocp and Odsa, and the possibility of including an Aurora module in the project.

Implementing New System With HPCN and CXL

Allan and Olivier discussed the implementation of a new system, with a focus on the use of HPCN and the potential for building a 4-module box. They also discussed the concept of CXL, which stands for Compute Express Link, and its potential applications. Geoffrey provided guidance on how to fill out a cost proposal template, emphasizing the importance of breaking down milestones and adjusting hours accordingly. Allan requested a copy of the template from Geoffrey, and they agreed to continue their discussions via email.