

Auto sizing of system reserved

Notes from Sig Node 01/12/2021:

- Maybe this is formula based limits
 - CPU
 - Memory
 - PID?
 - Ephemeral?
- Look more into PodOverhead

Motivation

Kubelet's *System Reserved* plays a crucial role in the OOMKilling the resource intensive pods. Without an adequate enough *System Reserved* we risk freezing the node making it completely unavailable for other pods.

In case of memory, we have observed that varying the value of *System Reserved* with respect to the installed capacity of the node helps to deduce an optimal value for it.

Currently, the only way to customize the System Reserved limits is to pre-calculate the values prior to Kubelet start. If the Kubelet is deployed to various instance types, then the limits need to be tuned for every instance type.

Proposal

Enable Kubelet to determine the value of the *System Reserved* automatically during start up.

Prior Work

Kubelet can refer to a pre-existing formula and table for guidance. E.g.

https://cloud.google.com/kubernetes-engine/docs/concepts/cluster-architecture#memory_cpu or

Elana Hashman has done some research in this area while working on Azure, and has come up with the following table that we find is very useful.

Memory (GB)	8	16	32	64	128	256
Reserved (GB)	1.8	2.6	3.56	5.48	9.32	11.88

Default Formula

Reserve an additional 100 MiB of memory on each node for kubelet eviction.

Allocatable resources are calculated in the following way:

$\text{Allocatable} = \text{Capacity} - \text{Reserved} - \text{Eviction Threshold}$

For memory resources, reserve the following:

- 255 MiB of memory for machines with less than 1 GB of memory
- 25% of the first 4GB of memory
- 20% of the next 4GB of memory (up to 8GB)
- 10% of the next 8GB of memory (up to 16GB)
- 6% of the next 112GB of memory (up to 128GB)
- 2% of any memory above 128GB

For CPU resources, reserve the following:

- 6% of the first core
- 1% of the next core (up to 2 cores)
- 0.5% of the next 2 cores (up to 4 cores)
- 0.25% of any cores above 4 cores

Design

New Kubelet Configuration Options:

EnableAutoSystemReserve: true
AutoSystemReserveProfile<Optional>: not default

Kubelet Startup:

- Enumerate the Memory and CPU on the Node
- Calculate the system and memory reserved using the default formula
- Log the calculations

Alpha:

- EnableAutoSystemReserve and AutoSystemReserveProfile functionality

Beta:

- Provide Custom Profiles from the configuration file?