Contribution ID: **66** Type: **not specified**

Heterogeneous Memory Management

Tuesday, November 13, 2018 2:45 PM (45 minutes)

Heterogeneous computing use massively parallel devices, such as GPU, to crunch through huge data-set. This talks intends to present the issues, challenges and problems related to memory management and heterogeneous computing. Issues and problems from one address space per device which makes exchanging or sharing data-set between devices and CPUs hard, complex and error prone.

Solutions involve a unified address space between devices and CPU often call SVM (Share Virtual Memory) or SVA (Share Virtual Address). In those unified address space a virtual address valid on CPUs is also valid on the devices. Talk will address both hardware and software solutions to this problem. Moreover it will consider ways to preserve the ability to use the device memory in those scheme.

Ultimately this talks is an opportunity to discuss memory placement, like for NUMA architecture, in a world where we not only have to worry about CPU but also about devices like GPU and their associated memory.

If it were not enough, we now also have to worry about memory hierarchy for each CPU or device. Memory hierarchy going from fast High Bandwidth Memory (HBM) to main memory (DDR DIMM) which can be order of magnitude slower, and finally to persistent memory which is large in size but slower and with higher latency.

I agree to abide by the anti-harassment policy

Yes

Primary author: GLISSE, Jerome (Red Hat)

Presenter: GLISSE, Jerome (Red Hat)

Session Classification: LPC Main Track

Track Classification: Refereed talk