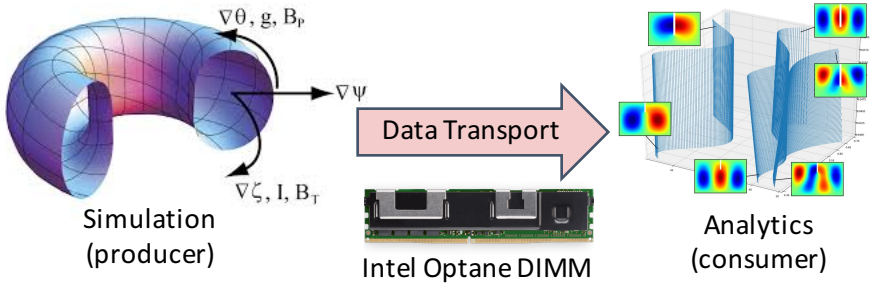
Leveraging Intel Optane for HPC Workflows

Ranjan Sarpangala Venkatesh[1], Tony Mason[1,2], Pradeep Fernando[1], Greg Eisenhauer[1], Ada Gavrilovska[1]
[1]Georgia Institute of Technology, [2]University of British Columbia

Background: Data movement in HPC workflows

- Pipeline of simulation and analytics apps
- Periodic compute and I/O phases
- Process large volumes of data
- ***In situ* execution** is widely used to reduce cross-node traffic



Problem

How to maximize the benefit that in situ workflows can obtain from using PMEM for their data exchanges?

Workflow deployment considerations

- Locality of PMEM access
  - Local Write (*Loc-W*) vs Local Read (*Loc-R*)
- Contention of PMEM resources
  - Serial (*S*) vs Parallel (*P*) execution

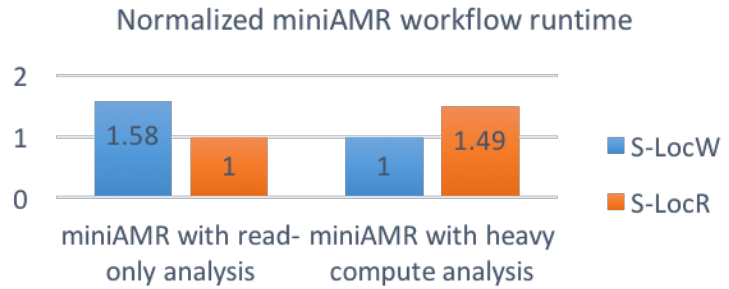
Workflow Streaming I/O parameters

- Iteration cycle composition of simulation and analytics
  - Lengths of compute and I/O phases
- I/O granularity – small vs large objects
- I/O concurrency - # MPI ranks, low, medium and high

Sources:
http://phoenix.ps.uci.edu/GTC/presentations/tutorial/GTC-01.pdf
https://www.epfl.ch/labs/csqi/
https://www.intel.com/

Challenge

- Optimizing individual workflow components does not guarantee an optimal choice of the end-to-end workflow performance.



Observations

Workflow characteristics	B/W utilization	Concurrency	Config
I/O intensive simulation and analytics kernels	High	Medium/High	*S-LocW*
Long compute simulation or use of small objects	Low	High	S-LocR
I/O intensive simulation with compute intensive analytics	Low	Low	*P-LocW*
Long compute simulation or use of small objects	Low	Low/Medium	*P-LocR*

Takeaway: Our results show naive use of PMEM degrades runtime up to 70%. Future HPC workflow schedulers have to consider compute and I/O characteristics of workflow components to make intelligent PMEM-aware choices in placement and execution decisions.