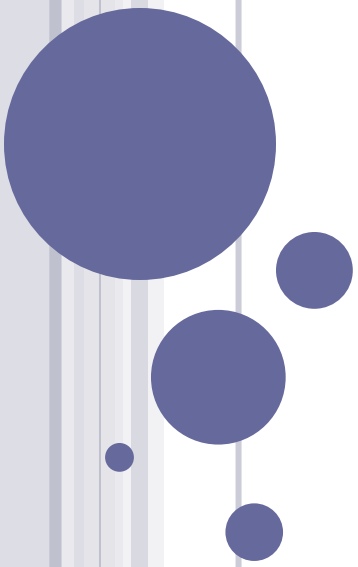


Update on DOE HPC Efforts

Matthew Wolf

CERCS Electronic IAB, Fall 2013

On-Going Research

- A number of existing projects and extensions on previous projects in the HPC space, particularly leading towards Exascale systems design
  - Transactional support for distributed workflows
  - Staging transports for ADIOS
  - Streaming paradigms for exascale runtimes
  - Planning and placement for co-scheduled execution & NVRAM execution
  - Containerized execution and management of HPC workflows
- Primary contacts:
  - Oak Ridge National Laboratory, Sandia National Laboratories (both sites)
- Associates:
  - Lawrence Berkeley National Laboratory, Intel, HP, UT Austin

Specific Projects

- These topics all come from particular joint funded projects:
  - Sandia: HPC internal funding
  - Sandia et al: Hobbes ExaOS proposal (new start)
  - LBNL et al: Scientific Data Analysis and Visualization (SDAV) Institute
  - ORNL et al: RSVP runtime staging (extension)
  - Sandia et al: ExaCT CoDesign Center
- Related:
  - NVM work w/Intel & HP, RDAV funding from NSF

Technical Components

- There are a couple of key threads that follow through multiple of the engagements

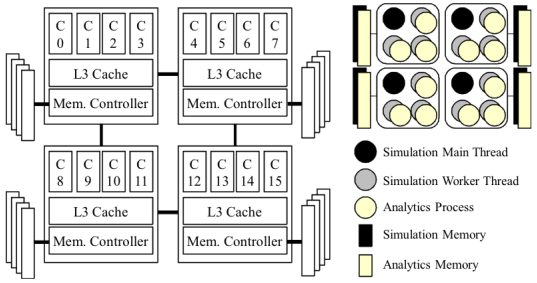
- Careful Data Movement
  - A lot has been invested recently in the integration effort with Sandia's NNTI transport level, which gives us an easy integration point with a large number of RDMA transports
  - This extension has also allowed us to start in-lining the lock-free shared memory implementations first demonstrated in the GoldRush (Fang Zheng et al) paper
  - Transactional support (Jai Dayal & Jay Lofstead) for tracking data and computation through in-memory workflows has also been a key component

Staging

- FlexPath innovations – extending our ADIOS investment
  - FlexPath is our new software target for integrating all of the high performance messaging lessons that we've learned from the enterprise world into the HPC streaming data environment
  - Inherits NNTI, sockets, & shared memory from EVPath
  - Extensions
    - Dynamic overlays for streaming workflow support (Yanwei Zhang)
    - Containerized replicas of workflow stages for throughput management (Jai Dayal)
    - Self-monitoring for dynamic reprovisioning (Xuechen Zhang)
- Because of these runtime improvements, staging is a programming style, not an assumption on resources
  - Could use resources on the source node, on some other nodes, or a distant machine

GoldRush

- Tie thread scheduling, NUMA memory placement, and in-situ analytics together



Placement of simulation and in situ data analytics on Smoky's 16-core compute node.

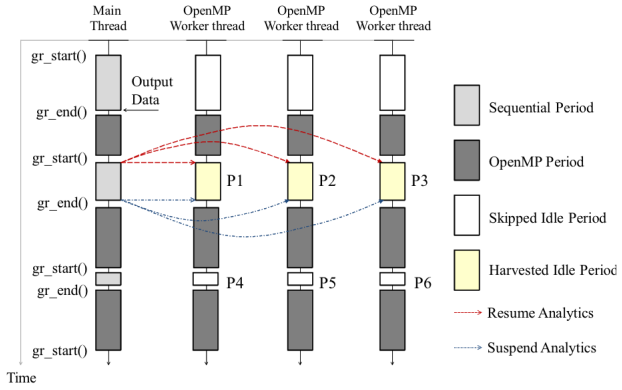


Figure 8. Simulation and analytics

- Paper shows a performance improvement by tweaking the OpenMP runtime to allow co-scheduling and placement of staging w/core algorithm.

OS & NVM

- Much of this is covered elsewhere
- Shortly – deep memory hierarchy and performance differences associated with it have now (re-)opened a lot of research topics
- Look for forthcoming discussions in this area!

Project Engagements

- In addition to the organizational contacts mentioned, there are numerous scientific engagements

- Astrophysics (Maya/Einstein's Toolkit)
- Fusion Science (XGC*, GTS, Pixie3D, Experimental)
- Materials Science (LAMMPS, new engagements)
- Combustion Science (S3D, Experimental)

- Analytics & Visualization (multiple)

Future directions

- More complex controls for staging/stream management
  - How to give users rational ways to reason about shared state, performance dependencies, etc.?
- Support for accelerators
  - On-going tie-in to heterogeneous multi-core support
  - Looking to engagements around language extensions (Intel)
- Stream programming models
  - Leveraging ADIOS platform, but looking at new languages bindings, support for scriptable interfaces, commodity/HPC interaction (html5, json, etc.)