**First Session: November 4, Tuesday 6-7 PM**

**Title:** Energy-efficiency and Resilience Challenges for Long-running Application on Leadership-scale Machines

**Organizer:** Devesh Tiwari (ORNL)

**Location:** Josephs Room

**Description:** Oak Ridge Leadership Computing Facility (OLCF) enables faster scientific discovery through its supercomputers (currently, Titan supercomputer ranked at no.2 in the Top 500 supercomputer list). OLCF has helped scientists analyze scientific problems at a scale that was not even imaginable a few years ago. But, we are facing new challenges as we prepare for the exa-scale era. Energy-efficiency and resilience being the two major challenges of the lot. The purpose of this BoF is to learn, share and explore collaborative-opportunities for improving the energy-efficiency and resilience of long-running applications.

Through this BoF session, we aim to explore opportunity for joint activities to understand the computational, communication and I/O requirements of emerging large-scale applications, e.g. next-generation sequencing, blood-flow simulations, etc. How can we partition computer systems resources in an energy-efficient way? How can these application benefit from a supercomputer as opposed to small-scale clusters in the cloud? What are the challenges and trade-offs in scaling-out these applications, with respect to energy-efficiency and resilience? What can system researchers do to alleviate soft-errors and data-movement cost in the supercomputer setting? So, bring on your favorite application. You are also welcome to present a 5 min lightning talk about your potential use-case or application requirements. Hopefully, this BoF and follow-on community activities will fill the gap between traditional systems research and the HPC challenges.

**Summary**

The title of this BOF is “Energy-efficiency and Resilience Challenges for Long-running Application on Leadership-scale Machines”. This BOF comprises two parts. The first part is a present from the organizer Devesh Tiwari. The second part is discussion led by the organizer. This report first summarizes the presentation and then highlights issues discussed in this session.

Devesh Tiwari first introduced by Oak Ridge Leadership Computing Facility (OLCF). It has Titan supercomputer, which has 18688 compute nodes and a 27 petaFlops peak performance (24.5 petaFlops from GPU and 2.5 petaFlops from CPU). Applications running in Titan are mostly very long-running, MPI-based, and scalable scientific applications, e.g. Molecular dynamics, power grid, energy storage, combustion, nuclear energy. Around 50% of its users are from domains of physics and chemistry. The mission of OLCF is to use computation resources at these facilities to do scientific discovery.

For challenges at OLCF, Devesh Tiwari first talked about power challenges. At OLCF, people are facing tremendous challenges on power. Comparing with its previous supercomputer Jaguar, Titan has improved the performance from 2.4 petaFlops to 27 petaFlops. While, Titan’s power consumption is 9 MW which is 2 MW higher than that of Jaguar. Although Titan achieves about 10 times performance improvement with a 28% increase on power consumption. This trend is not enough to achieve the goal which is 1 exaFlops at 20 MW, a 500 times performance improvement with a 2 times increase on power consumption. Although GPU is energy efficient, it is still not enough to achieve the goal. Based on observations, data-movement cost is likely to dominate power consumption. Also, new memory techniques are not enough to reduce the power consumption. People need not just think how to scale and also need to to rethink how to manage flops and memory accesses with respect to energy cost.

Then, Devesh Tiwari talked about resilience challenges. At OLCF, scientific applications are long-running applications which use checkpointing as a mechanism to recover from failures. In the presence of failure, an application restarts from its last available checkpoint. Work made during the last available checkpoint and the point of failure is waste work. The frequency of checkpointing depends both on application specifications and resource specifications (like the I/O bandwidth for checkpointing). Devesh Tiwari showed a figure in which, for a 4-hour checkpointing interval and a given application, the portion of execution time contributed by waste time increased significantly with the increase of the number nodes used in this application. Based on this trend, at exascale, checkpointing/restart overhead may be higher than the time of computing itself. Also, at exascale, a failure is expected every hour in a room full of 100K nodes. Checkpointing cost would be of the same order of computing time. The question is if we can still use traditional checkpoint/restart techniques. Devesh Tiwari also mentioned that soft-errors pose a bigger challenge. When a soft-error happens, it will be great if the application crashes. The challenging issues are to deal with silent data corruption and incorrect results which cannot be verified.

In the presentation, Devesh Tiwari listed several questions which do not have a good answer yet. Those are summarized below.

Questions on power consumption:

* What are the compute and I/O requirements of emerging leadership applications? How to do job scheduling and Resource provisioning?
* Can users manage energy as a resource, like cache, GPUs?
	+ Devesh Tiwari thought we can. He also mentioned that if we can have energy-efficiency APIs or programmer hints.
* Should we establish an energy usage based pricing model? If we should, how to account energy? Will users trust our energy accounting?
	+ Current pricing model: node hours.
* How to collect and share energy accounting data for different components, e.g. compute, network and storage?
* How to trade-off data-movement and performance?
	+ Re-computing vs. accessing memory?
	+ Communication-avoiding algorithm?

Questions on resilience:

* Should we get rid of checkpoint/restart mechanism?
	+ Algorithm based on fault-tolerance
	+ Reversible computing
	+ N-modular redundancy
* Dealing with silent data corruption
	+ Need for systematic methodology and field study
	+ Which algorithms can tolerate such errors
	+ Data corruption can happen  inside processor cores
* Architecture specific studies (GPU vs. different CPU generations)
* Application specific resilience mechanism
	+ Which part of the data needs more protection?
	+ Programming APIs for resilience?
* Architecture specific studies (GPU vs. different CPU generations)
* Application specific resilience mechanism
	+ Which part of the data needs more protection?
	+ Programming APIs for resilience?

After the presentation from Devesh Tiwari, people had discussions on those challenges on power and resilience. We are highlighting their discussions below.

* Measuring Energy in OS.
	+ Not mature yet. Need to access sensors from OS (no access right now) and information is coarse-grained.
* Virtualization
	+ No virtualization technique is used in in supercomputers.
* Penalizing users who use over the budge of energy?
* Devesh Tiwari said not to suggest a new computing model.
* Cannot reduce power consumption of communication links. People expect the communication cost on power may be equal to that on computing.
* Communication is the problem for scientific computing in cloud. Communication links in cloud are slow.
* Locality-aware scheduling methods can solve some energy problems.