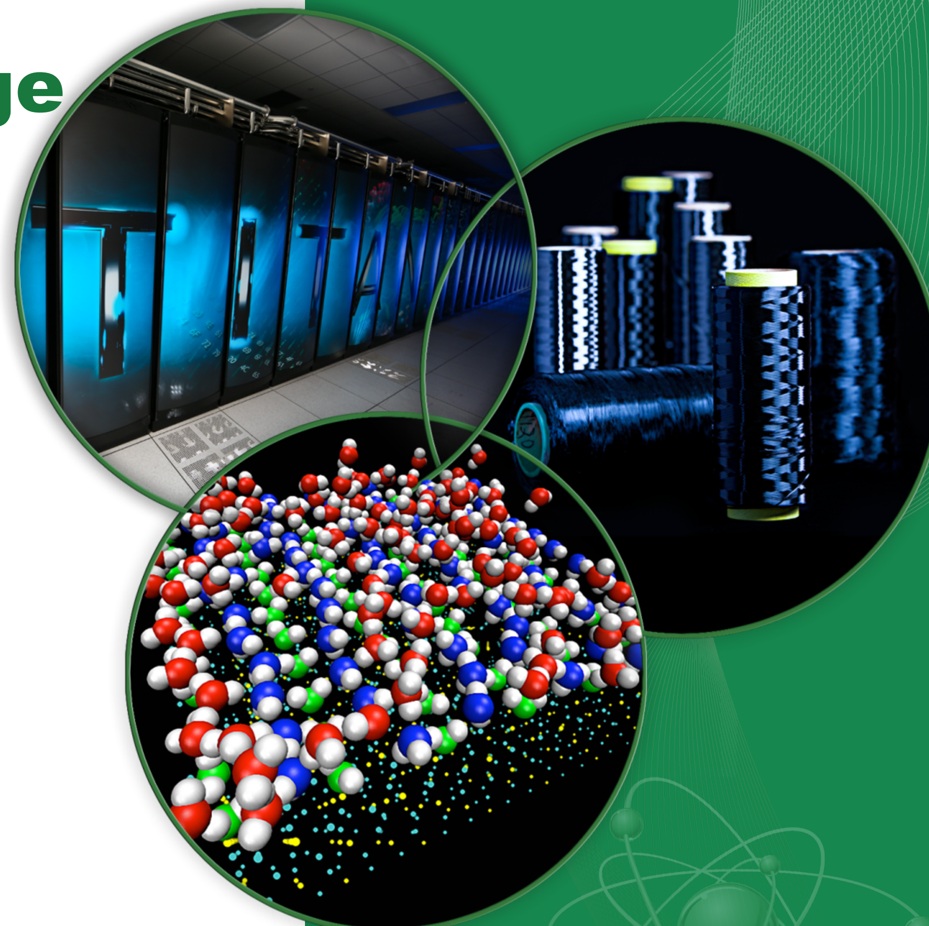
Measuring GPU Usage on Cray XK7 using NVIDIA's NVML and Cray's RUR



Jim Rogers

Director of Operations
National Center for Computational Sciences
Oak Ridge National Laboratory

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

CUG 2014 Technical Session 12B

Measuring GPU Usage on Cray XK7 using NVIDIA's NVML and Cray's RUR

ORNL introduced a 27PF Cray XK7 in to production in May 2013. This system provides users with 18,688 hybrid compute nodes, where each node couples an AMD 6274 Opteron with an NVIDIA GK110 (Kepler) GPU. Beginning with Cray's OS version CLE 4.2UP02, new features available in the GK110 device driver, the NVIDIA Management Library, and Cray's Resource Utilization software provide a mechanism for measuring GPU *usage* by applications on a per-job basis. By coupling this data with job data from the workload manager, fine grained analysis of the use of GPUs, by application, are possible. This method will supplement, and eventually supplant an existing method for identifying GPU-enabled applications that detects, at link time, the libraries required by the resulting binary (ALTD, the Automatic Library Tracking Database). Analysis of the new mechanism for calculating per-application GPU usage is provided as well as results for a range of GPU-enabled application codes.

Presenter Overview

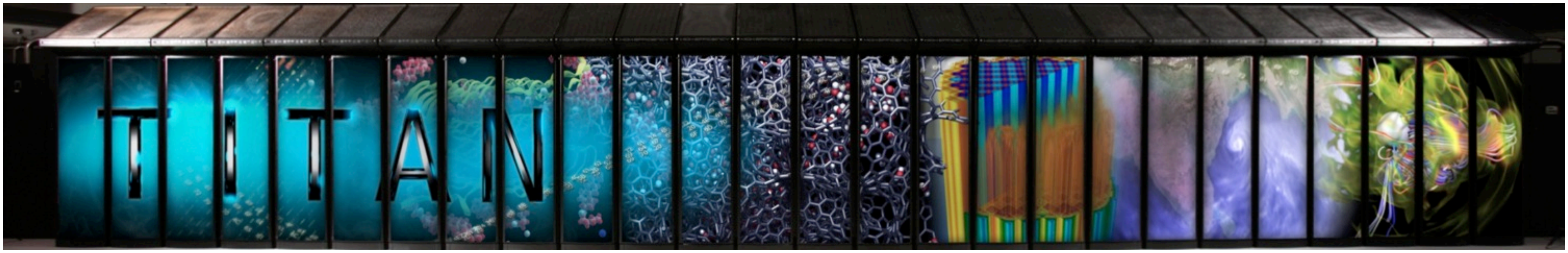
Jim Rogers is the Director of Operations for the National Center for Computational Sciences at Oak Ridge National Laboratory. The NCCS provides full facility and operations support for three petaFLOP-scale systems including Titan, a 27PF Cray XK7. Jim has a BS in Computer Engineering, and has worked in high performance computing systems acquisition, integration, and operation for more than 25 years.

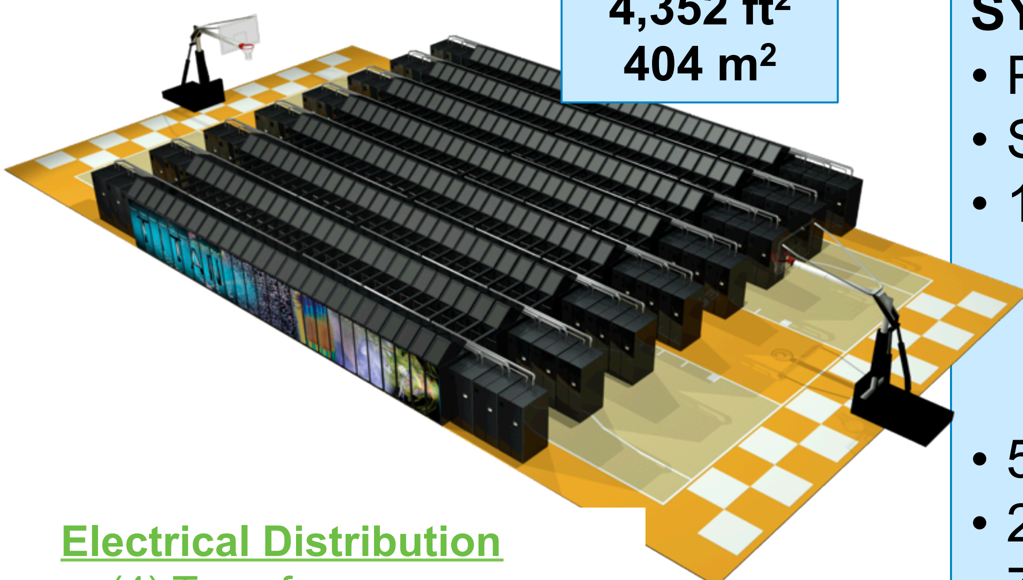
Content

- The OLCF's Cray XK7 Titan
  - Hardware Description
  - Assessing the Operational Impact to Delivered Science
    - Time- and Energy- to Solution. Case Study: WL-LSMS
- The Operational Need to Understand Usage
  - ALTD (the early years)
  - NVIDIA's Role
    - Δ to the Kepler Driver, API, and NVML
  - Cray's Resource Utilization (RUR)
- Examples of NVML_COMPUTEMODE_EXCLUSIVE_PROCESS Measurement
  - Lattice QCD
  - LAMMPS
  - NAMD
- Next Steps…
- INCITE Allocation Program

ORNL's Cray XK7 Titan

A Hybrid System with 1:1 AMD Opteron CPU and NVIDIA Kepler GPU





4,352 ft^2
404 m^2

Electrical Distribution

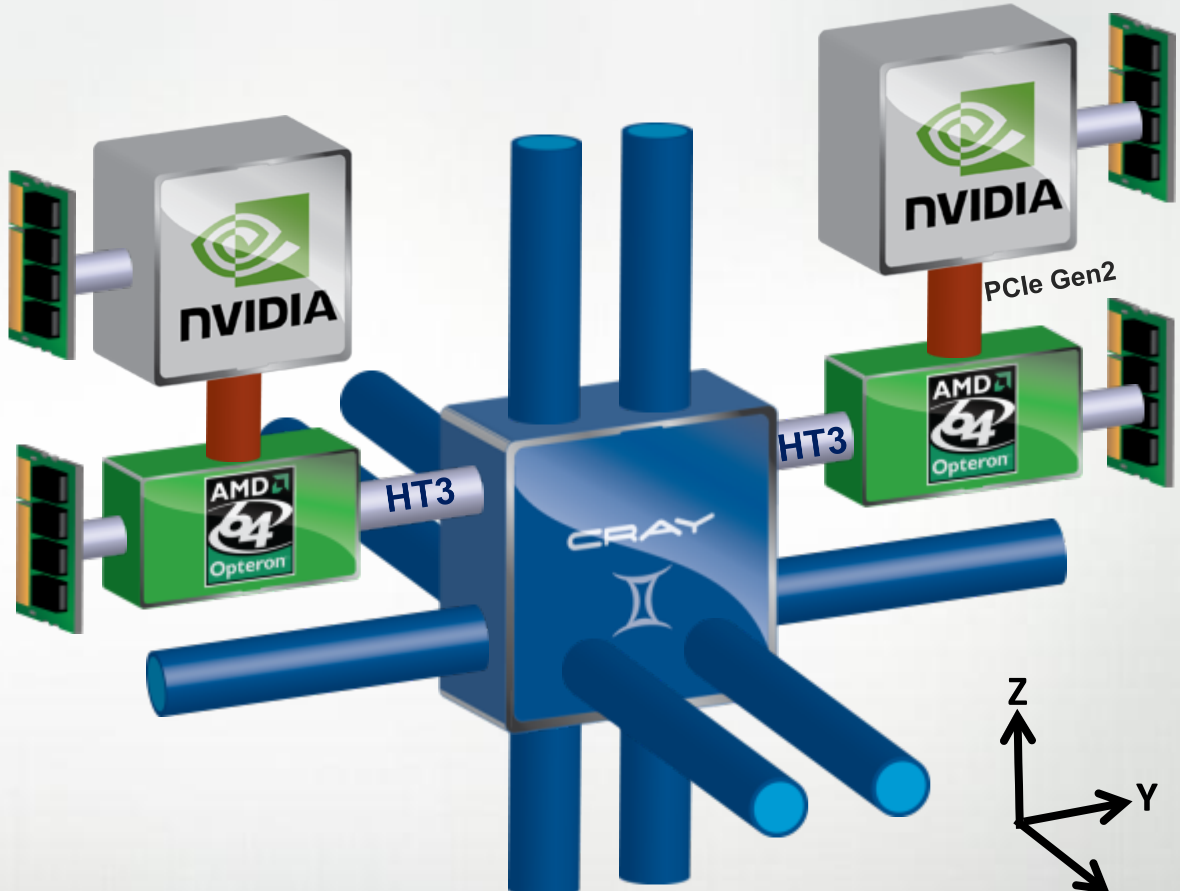
- (4) Transformers
- (200) 480V/100A circuits
- (48) 480V/20A circuits

SYSTEM SPECIFICATIONS:

- Peak performance of 27 PF
- Sustained performance of 17.59 PF
- 18,688 Compute Nodes each with:
    - 16-Core AMD Opteron CPU
    - NVIDIA K20x (Kepler) GPU
    - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 8.9 MW peak energy measurement

Cray XK7 Compute Node

XK7 Compute Node Characteristics
AMD Opteron 6274 16 core processor - 141 GF
Tesla K20x - 1311 GF
Host Memory 32GB 1600 MHz DDR3
Tesla K20x Memory 6GB GDDR5
Gemini High Speed Interconnect

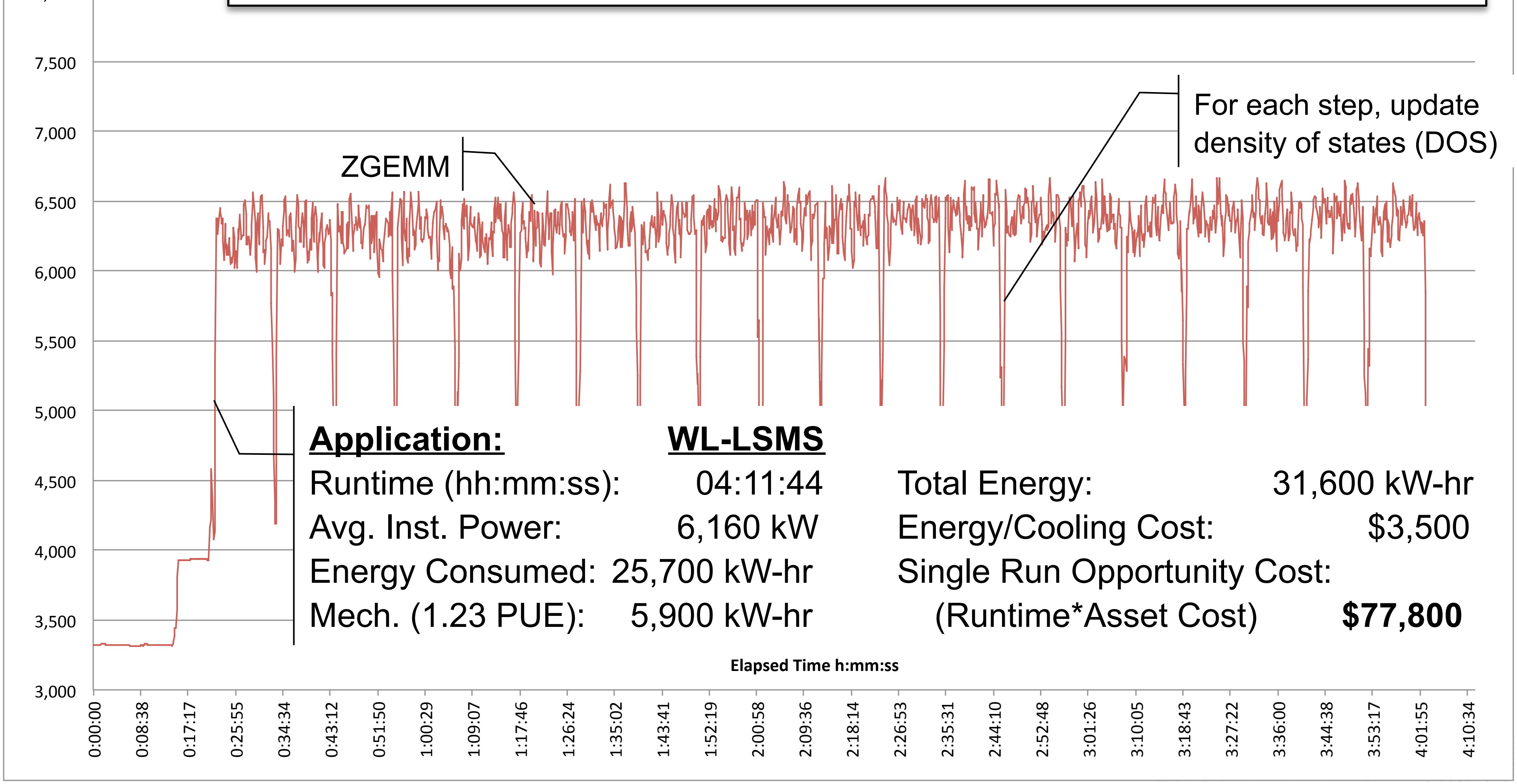


Slide courtesy of Cray, Inc.

Application Power Efficiency on the Cray XK7

The Behavior of Magnetic Systems with WL-LSMS

CPU-only power (energy) consumption trace for a WL-LSMS run that simulates 1024 Fe atoms as they reach their Curie temperature. Run size: 18,561 Titan nodes (99% of Titan). Run signature: Initialization, followed by 20 Monte Carlo steps. Computation dominated by double complex matrix matrix multiplication.

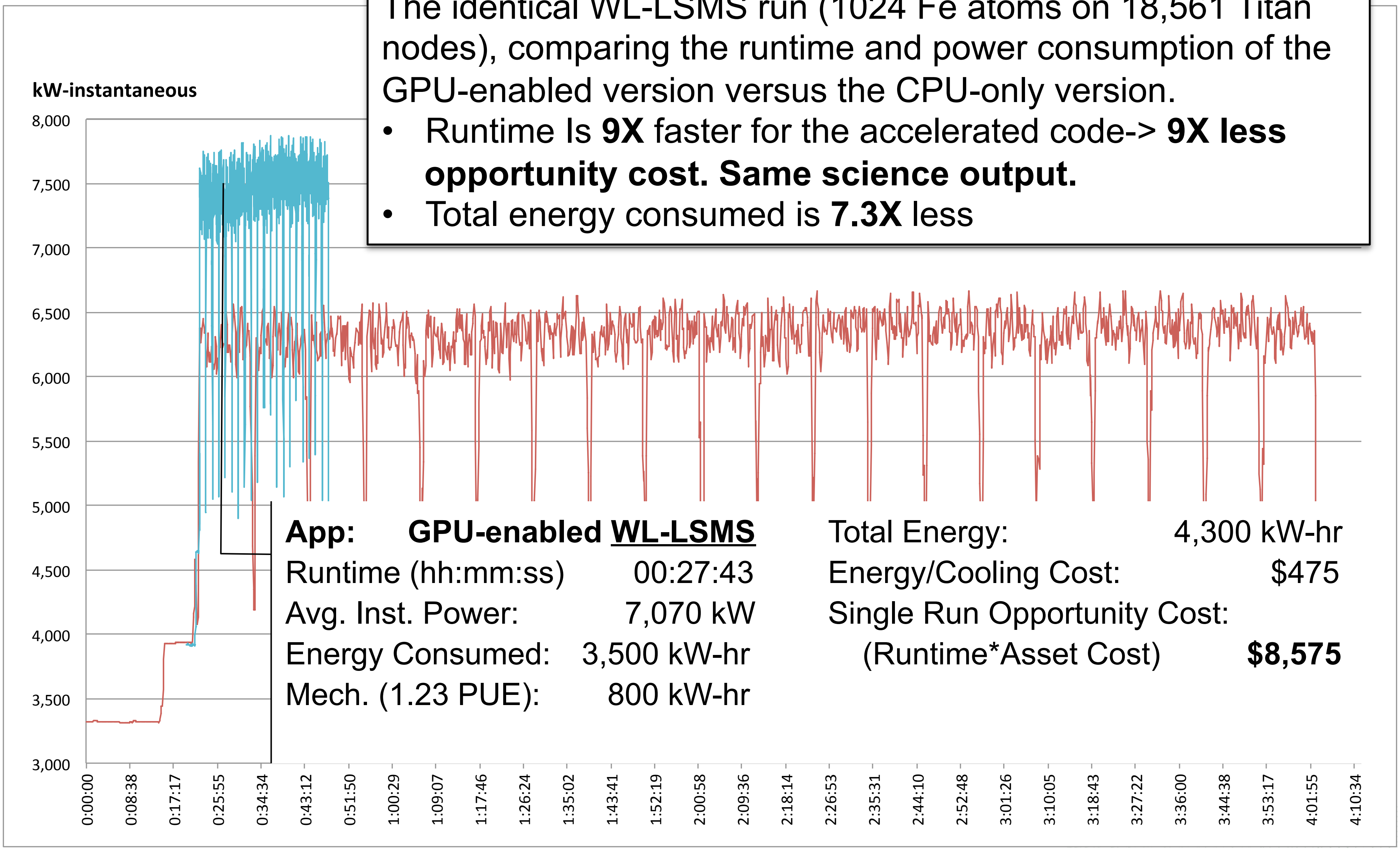


Application Power Efficiency on the Cray XK7

Comparing CPU-Only and GPU-Enabled WL-LSMS

The identical WL-LSMS run (1024 Fe atoms on 18,561 Titan nodes), comparing the runtime and power consumption of the GPU-enabled version versus the CPU-only version.

- Runtime Is **9X** faster for the accelerated code-> **9X less opportunity cost. Same science output.**
- Total energy consumed is **7.3X** less



App:	**GPU-enabled WL-LSMS**
Runtime (hh:mm:ss)	00:27:43
Avg. Inst. Power:	7,070 kW
Energy Consumed:	3,500 kW-hr
Mech. (1.23 PUE):	800 kW-hr

Total Energy:	4,300 kW-hr
Energy/Cooling Cost:	$475
Single Run Opportunity Cost: (Runtime*Asset Cost)	**$8,575**

Content

- The OLCF's Cray XK7 Titan
  - Hardware Description
  - Assessing the Operational Impact to Delivered Science
    - Time- and Energy- to Solution. Case Study: WL-LSMS
- The Operational Need to Understand Usage
  - ALTD (the early years)
  - NVIDIA's Role
    - Δ to the Kepler Driver, API, and NVML
  - Cray's Resource Utilization (RUR)
- Examples of NVML_COMPUTEMODE_EXCLUSIVE_PROCESS Measurement
  - Lattice QCD
  - LAMMPS
  - NAMD
- Next Steps…
- INCITE Allocation Program

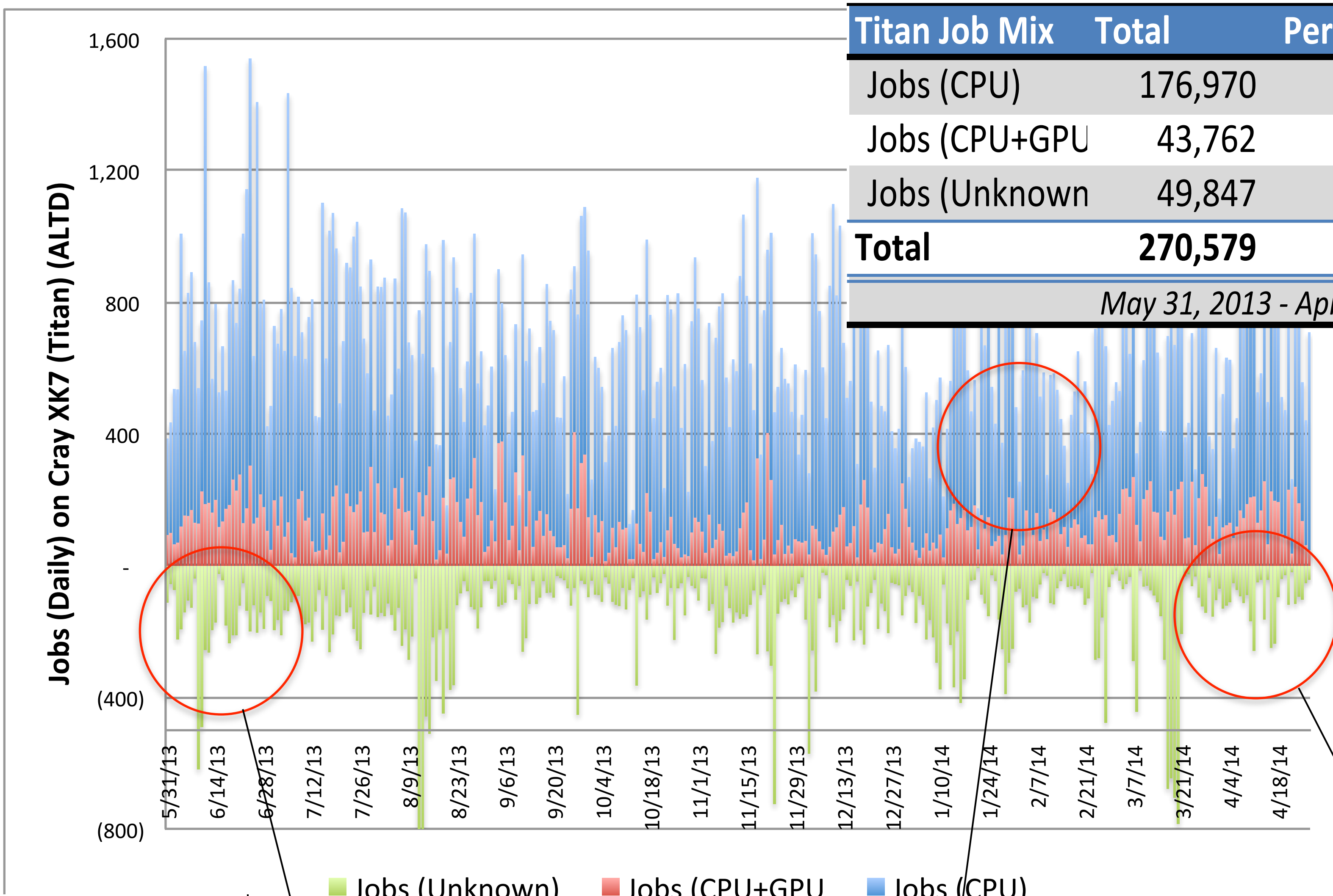
Monitoring GPU Usage on Titan- The Early Years

- Requirement: Detect, on a per-job basis, if/ when jobs use accelerator-equipped nodes.
- Initial Solution
  - Leverage ORNL's Automatic Library Tracking Database (ALTD)
    - At link time, a list of libraries linked against is stored in a database
    - When the resulting program is executed via aprun, a new ALTD record is written that contains the specific executable, to be run, the batch job id, and other info
  - Batch jobs are compared against ALTD to see if they were linked against an accelerator-specific library
    - libacc*, libOpenCL*, libmagma*, libhmpp*, libcuda*, libcupti*, libcula*, libcublas*
    - Jobs whose executables are linked against one of the above are deemed to have used the accelerator
- Outliers
  - Job run outside of the batch system
    - ALTD knows about them, but we can't tie them to usage because there's no job record
  - ALTD is enabled by default, but if it's disabled we won't capture link/run info

Making sense of an example link statement

% lsms /usr/lib/../lib64/crt1.o /usr/lib/../lib64/crti.o
/opt/gcc/4.7.0/snos/lib/gcc/x86_64-suse-linux/4.7.0/crtbegin.o
libLSMS.aSystemParameters.o libLSMS.aread_input.o
libLSMS.aPotentialIO.o
libLSMS.abuildLIZandCommLists.o
libLSMS.aenergyContourIntegration.o
libLSMS.asolveSingleScatterers.o libLSMS.acalculateDensities.o
libLSMS.acalculateChemPot.o
/lustre/widow0/scratch/larkin/lsms3-trunk/lua/lib/liblua.a
…
-lcublas /opt/nvidia/cudatoolkit/5.0.28.101/lib64/libcublas.so -lcupti
/opt/nvidia/cudatoolkit/5.0.28.101/extras/CUPTI/lib64/libcupti.so -
lcudart
/opt/nvidia/cudatoolkit/5.0.28.101/lib64/libcudart.so -lcuda
/opt/cray/nvidia/default/lib64/libcuda.so
/opt/cray/atp/1.4.4/lib//libAtpSigHCommData.a -lAtpSigHandler
/opt/cray/atp/1.4.4/lib//libAtpSigHandler.so -lgfortran
/opt/gcc/4.7.0/snos/lib/gcc/x86_64-suse-linux/4.7.0/../../../../lib64/
libgfortran.so -lhdf5_hl_cpp_gnu
...
/opt/cray/pmi/3.0.1-1.0000.9101.2.26.gem/lib64/libpmi.so -lalpslli
/usr/lib/alps/libalpslli.so -lalpsutil /usr/lib/alps/libalpsutil.so
/lib64/libpthread.so.0 -lstdc++
/lib64/ld-linux-x86-64.so.2 -lgcc_s
/opt/gcc/4.7.0/snos/lib/gcc/x86_64-suse-linux/4.7.0/../../../../lib64/
libgcc_s.so /opt/gcc/4.7.0/snos/lib/gcc/x86_64-suse-linux/4.7.0/
crtend.o
/usr/lib/../lib64/crtn.o

Assessing GPU Usage with ALTD – Job Distribution



Titan Job Mix	Total	Percentage
Jobs (CPU)	176,970	65.4%
Jobs (CPU+GPU	43,762	16.2%
Jobs (Unknown	49,847	18.4%
Total	**270,579**	**100.0%**
		May 31, 2013 - April 27, 2014

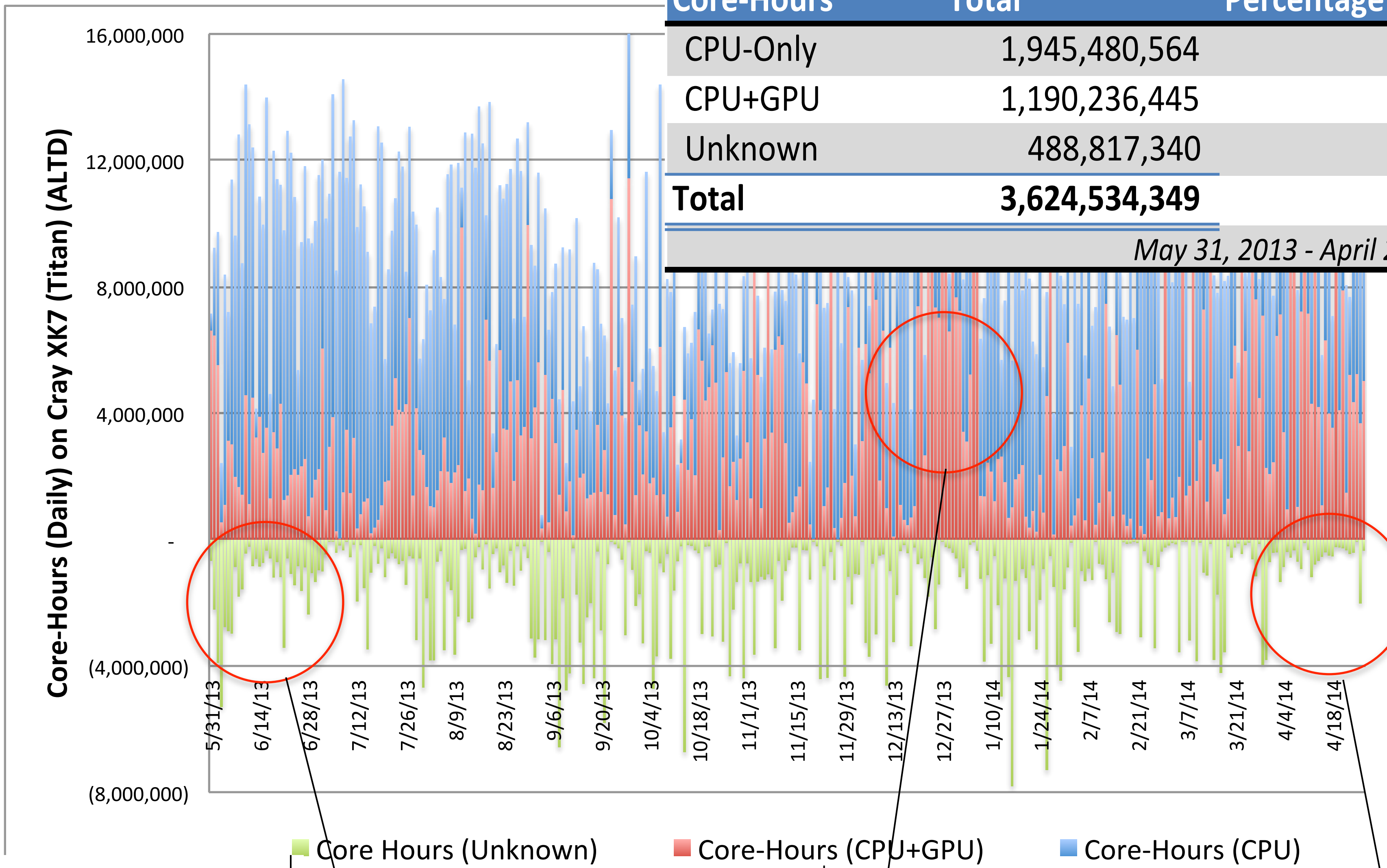
Rocky start using ALTD… lots of edge cases escaped.

Danger- Measuring job *counts* does not account for the work produced by an individual job

Unknowns are 18.4% of total delivered *jobs* since May 31, 2013.

Assessing GPU Usage with ALTD – Core Hours

Core-Hours	Total	Percentage
CPU-Only	1,945,480,564	53.7%
CPU+GPU	1,190,236,445	32.8%
Unknown	488,817,340	13.5%
Total	**3,624,534,349**	100.0%
		May 31, 2013 - April 27, 2014



Rocky start using ALTD… lots of edge cases escaped.

Great *apparent* use of the GPU by the workflow, but no way to quantify it.

Unknowns are 13.5% of total delivered hours since May 31, 2013.

NVIDIA's Role –
Δ to the Kepler Driver, API, and NVML

The previous NVML is cool. You can spot check…

- Driver version
- pstate
- Memory use
- Compute mode
- GPU utilization
- Temperature
- Power
- Clock

But we needed…

- GPU utility (not point in time utilization) for the life of a process
- Persistent state of that GPU and memory data.
- Ability to retrieve that data, by apid, using a predefined API

And we conceded…

- if there is work on any of the 14 SMs, we are accumulating GPU utility.

- NVIDIA products containing these new features
  - Kepler (GK110) or better;
  - Kepler Driver 319.82 or later;
  - NVML API 5.319.43 or later;
  - The CUDA 5.5 release cadence

nvidia-smi Output (truncated) from a Single Titan Kepler GPU

```
==============NVSMI LOG==============

Timestamp                           : Mon Mar 18
16:51:15 2013
Driver Version                      : 304.47.13
Attached GPUs                       : 1
GPU 0000:02:00.0
    Product Name                    : Tesla K20X
    Display Mode                    : Disabled
    Persistence Mode                : Enabled
    Performance State               : P8
    Clocks Throttle Reasons
        Idle                        : Active
        User Defined Clocks         : Not Active
        SW Power Cap                : Not Active
        HW Slowdown                 : Not Active
        Unknown                     : Not Active
    Memory Usage
        Total                       : 5759 MB
        Used                        : 37 MB
        Free                        : 5722 MB
    Compute Mode                    :
Exclusive_Process

        Gpu                         : 0 %
        Memory                      : 0 %
    Ecc Mode
        Current                     : Enabled
        Pending                     : Enable
```

```
    Power Management                : Supported
    Power Draw                      : 18.08 W
                                      00 W
                                      00 W
                                      0 W
    Max Power Limit                 : 300.00 W
Clocks
                                      MHz
                                      MHz
                                      MHz
    Graphics                        : 732 MHz
    Memory                          : 2600 MHz
Max Clocks
                                      MHz
                                      MHz
    Memory                          : 2600 MHz
Compute Processes                   : None
```

Driver version for XK is no less than 304.47.13

Kepler – the K20X

Kepler has either a p-state of 0 (busy) or 8 (idle)

6GB GDDR5

GPU Utilization. HOWEVER- This is a point-in-time sample, and has no temporal quality.

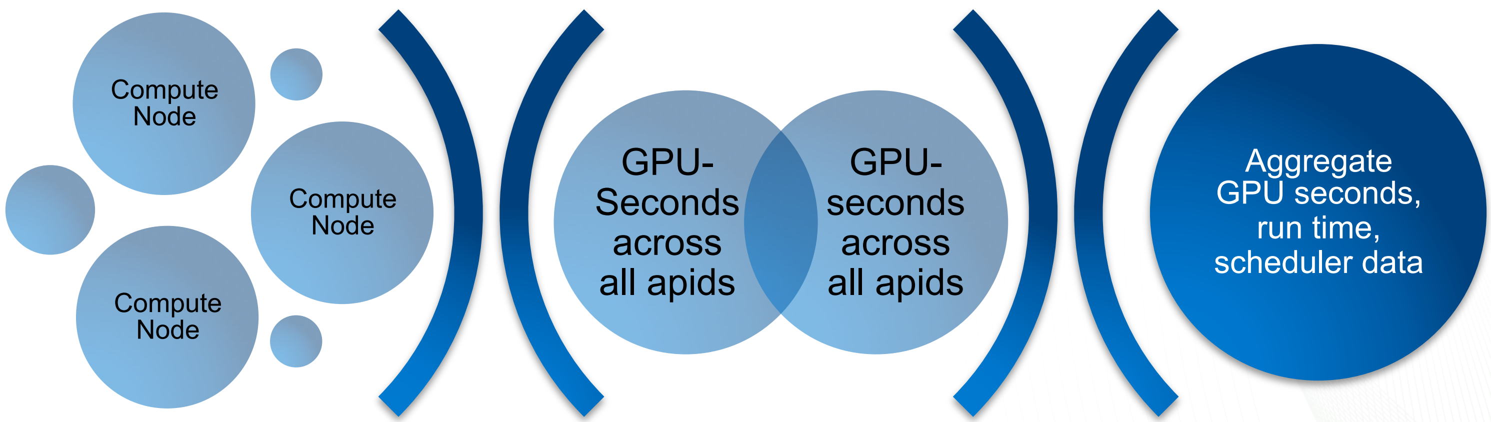
NVML is a C-based API for monitoring and managing various states of the NVIDIA GPU devices. ***nvidia-smi*** is an existing application that uses the nvml API.

Caution: Default Mode versus Exclusive Process

- The default GPU compute mode on Titan is EXCLUSIVE_PROCESS. However, we do not preclude users from using DEFAULT compute mode, and some applications demonstrate slightly better performance in DEFAULT compute mode.
- In EXCLUSIVE_PROCESS compute mode, the current release of the Kepler device driver acts exactly like you would expect.
- *However, in Default Mode, the aggregation of GPU seconds across multiple contexts can be misinterpreted by third party software using the new API.*
  - Look for updates to the way that GPU seconds are accumulated across multiple contexts in Default mode as the CUDA 6.5 cadence nears.
- Kepler Compute Modes:
  - NVML_COMPUTEMODE_DEFAULT Default compute mode – multiple contexts per device.
  - NVML_COMPUTEMODE_EXCLUSIVE_THREAD Compute-exclusive-thread mode – only one context per device, usable from one thread at a time.
  - NVML_COMPUTEMODE_PROHIBITED Compute-prohibited mode – no contexts per device.
  - NVML_COMPUTEMODE_EXCLUSIVE_PROCESS Compute-exclusive-process mode – only one context per device, usable from multiple threads at a time.

Cray RUR, and the NVIDIA API

- At the conclusion of every job, Cray uses the revised NVIDIA API to query every compute node associated with a job, extracting the accumulated GPU usage and memory usage statistics on each individual node.
- By aggregating that information with data from the job scheduler, statistics can then be generated that describe the GPU usage, on a per-job basis.



Collect GPU utility, per node, for a specific apid

Aggregate GPU Seconds (utility * runtime) across an entire aprun. Reference scheduler data for run time and other data

Provide specific job information that allows determination of GPU usage, per job

Extending the RUR Functionality…

Credit: Don Maxwell, Mitch Griffith, Adam Carlyle

- ORNL collects information from multiple sources, including the workload manager and Cray's RUR including:
  - Aprun ID,
  - Job Mode,
  - Start Time, End Time, Total Seconds,
  - Nodes, Total Node Seconds,
  - Tasks,
  - GPU Seconds,
  - Memory Usage
  - Command Name

- Allowing us to assemble reports that provide per-application granularity of:
  - Number of distinct runs of the same application (workflow/ production)
  - Percentage of time, per run, during which the GPU was active
  - Less granular variations on the theme (aggregate assessments across all delivered hours, etc)

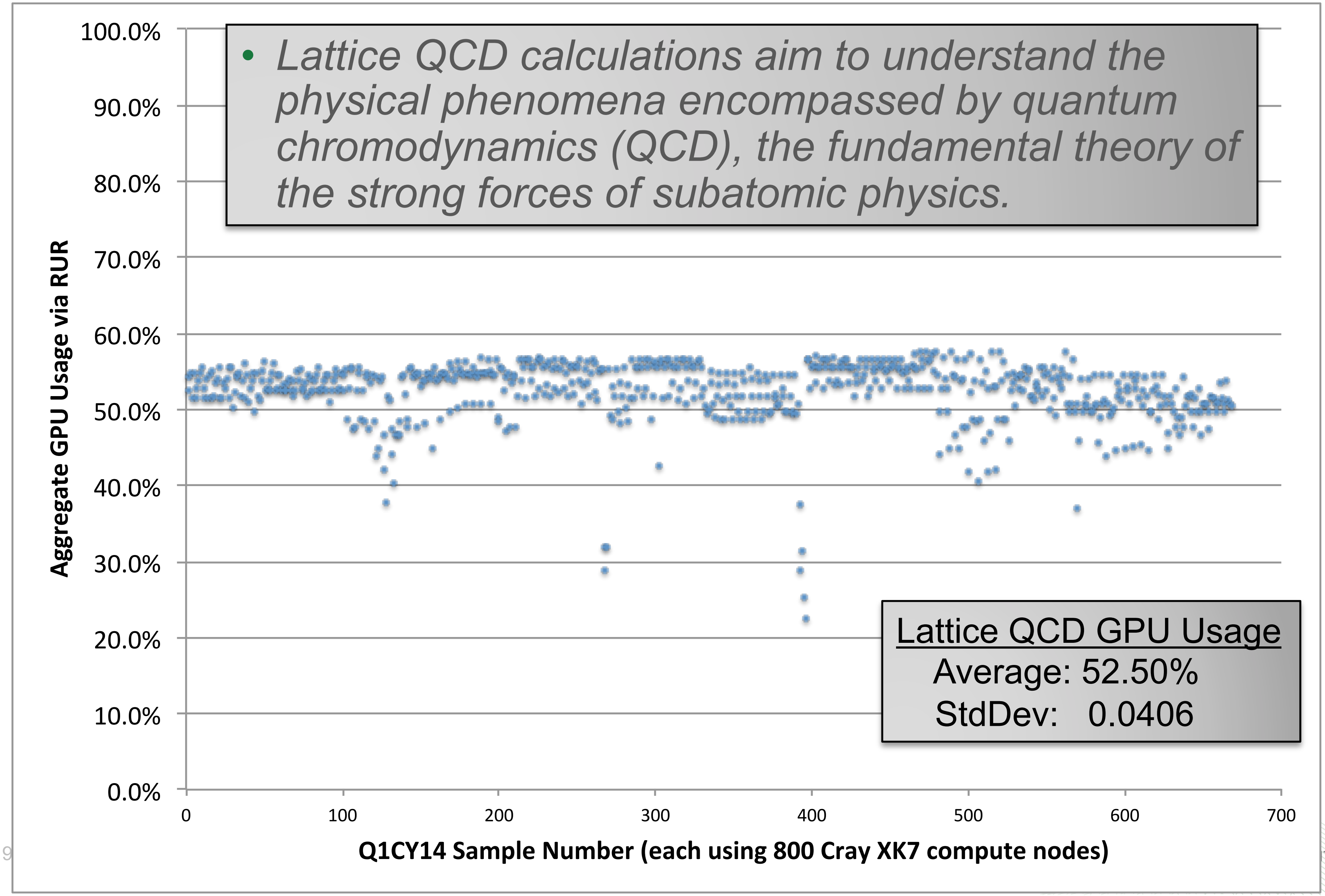
4370635	None	2014-01-29	2014-01-29	4124	800	3299200	1769720	12800	1,769,720	5408948224	4.33E+12	53.60%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370749	None	2014-01-29	2014-01-29	2444	800	1955200	1010471	12800	1,010,471	5408948224	4.33E+12	51.70%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370634	None	2014-01-29	2014-01-29	4112	800	3289600	1798702	12800	1,798,702	5408948224	4.33E+12	54.70%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370614	None	2014-01-29	2014-01-29	4271	800	3416800	1847710	12800	1,847,710	5408948224	4.33E+12	54.10%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370621	None	2014-01-29	2014-01-29	2413	800	1930400	1010934	12800	1,010,934	5408948224	4.33E+12	52.40%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370589	None	2014-01-29	2014-01-29	4195	800	3356000	1766529	12800	1,766,529	5408948224	4.33E+12	52.60%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370581	None	2014-01-29	2014-01-29	4164	800	3331200	1788014	12800	1,788,014	5408948224	4.33E+12	53.70%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370579	None	2014-01-29	2014-01-29	4177	800	3341600	1827159	12800	1,827,159	5408948224	4.33E+12	54.70%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370594	None	2014-01-29	2014-01-29	2405	800	1924000	1011823	12800	1,011,823	5408948224	4.33E+12	52.60%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370540	None	2014-01-29	2014-01-29	4162	800	3329600	1854590	12800	1,854,590	5408948224	4.33E+12	55.70%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370555	None	2014-01-29	2014-01-29	2404	800	1923200	1010791	12800	1,010,791	5408948224	4.33E+12	52.60%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370513	None	2014-01-29	2014-01-29	4224	800	3379200	1781072	12800	1,781,072	5408948224	4.33E+12	52.70%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370505	None	2014-01-29	2014-01-29	4099	800	3279200	1764247	12800	1,764,247	5408948224	4.33E+12	53.80%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370501	None	2014-01-29	2014-01-29	4062	800	3249600	1776503	12800	1,776,503	5408948224	4.33E+12	54.70%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370507	None	2014-01-29	2014-01-29	2411	800	1928800	1013741	12800	1,013,741	5408948224	4.33E+12	52.60%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370486	None	2014-01-29	2014-01-29	3891	800	3112800	1699424	12800	1,699,424	5408948224	4.33E+12	54.60%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc
4370356	None	2014-01-29	2014-01-29	4459	800	3567200	1821367	12800	1,821,367	5408948224	4.33E+12	51.10%	hmc	/lustre/atlas2/lgt003/scratch/bjoo/run/aniso/hmc

Content

- The OLCF's Cray XK7 Titan
  - Hardware Description
  - Assessing the Operational Impact to Delivered Science
    - Time- and Energy- to Solution. Case Study: WL-LSMS
- The Operational Need to Understand Usage
  - ALTD (the early years)
  - NVIDIA's Role
    - Δ to the Kepler Driver, API, and NVML
  - Cray's Resource Utilization (RUR)
- Examples of NVML_COMPUTEMODE_EXCLUSIVE_PROCESS Measurement
  - Lattice QCD
  - LAMMPS
  - NAMD
- Next Steps…
- INCITE Allocation Program

GPU Usage by Lattice QCD on OLCF's Cray XK7 Titan

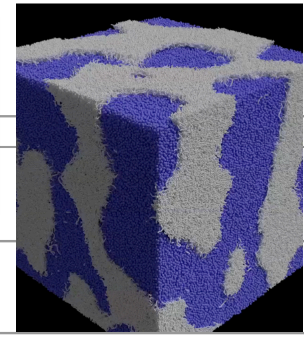
NVML_COMPUTEMODE_EXCLUSIVE_PROCESS



GPU Usage by LAMMPS on OLCF's Cray XK7 Titan

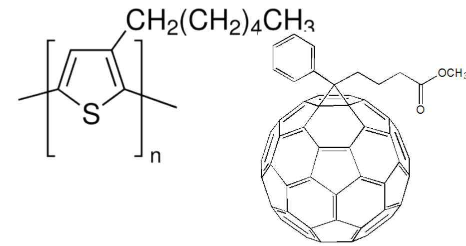
Mixed Mode (OpenMP + MPI), NVML_COMPUTEMODE_EXCLUSIVE_PROCESS

LAMMPS - Classical Molecular Dynamics Software used in simulations for biology, materials science, granular, mesoscale, etc

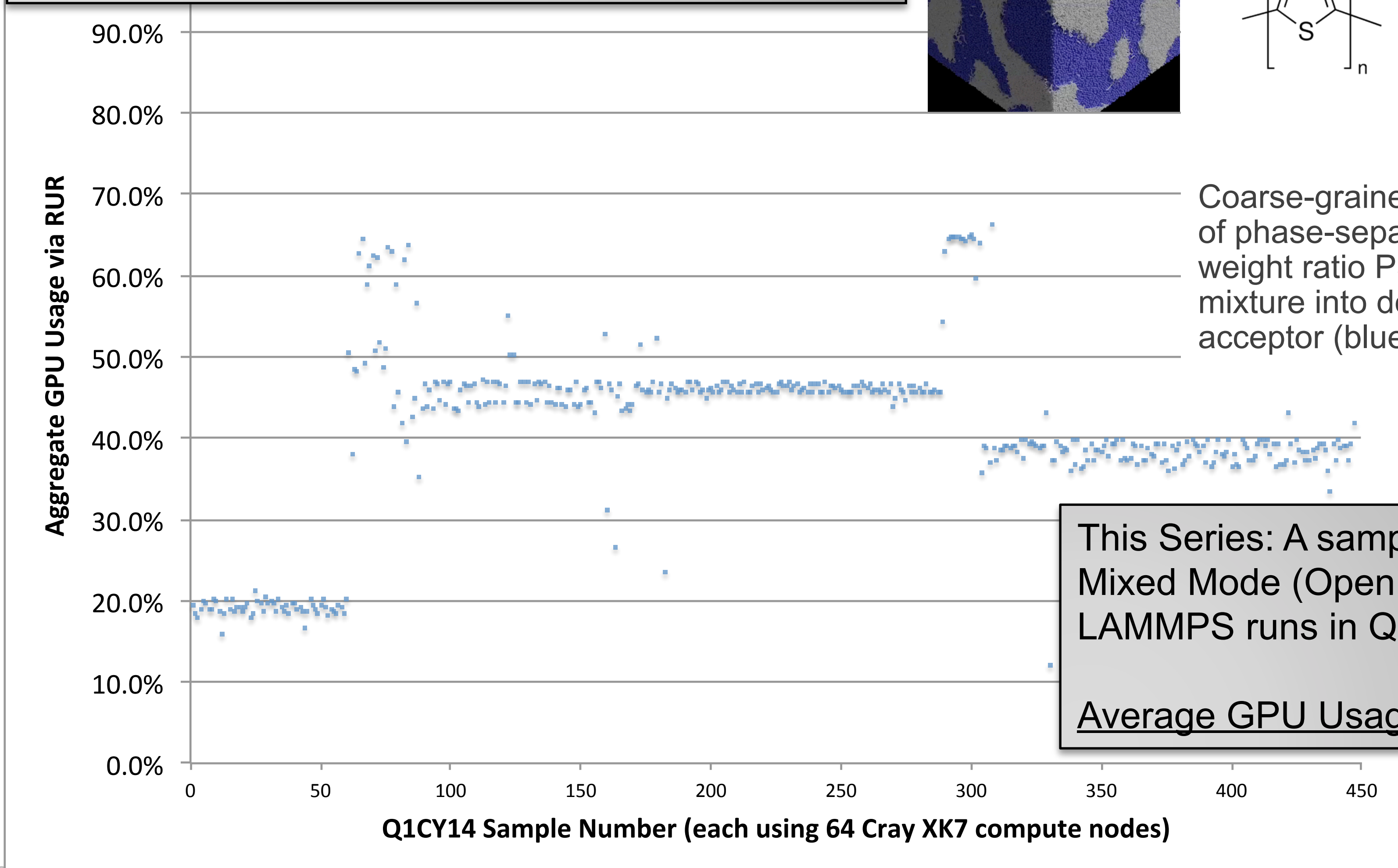


P3HT
(electron donor)

PCBM
(electron acceptor)



Coarse-grained MD simulation of phase-separation of a 1:1 weight ratio P3HT/PCBM mixture into donor (white) and acceptor (blue) domains.



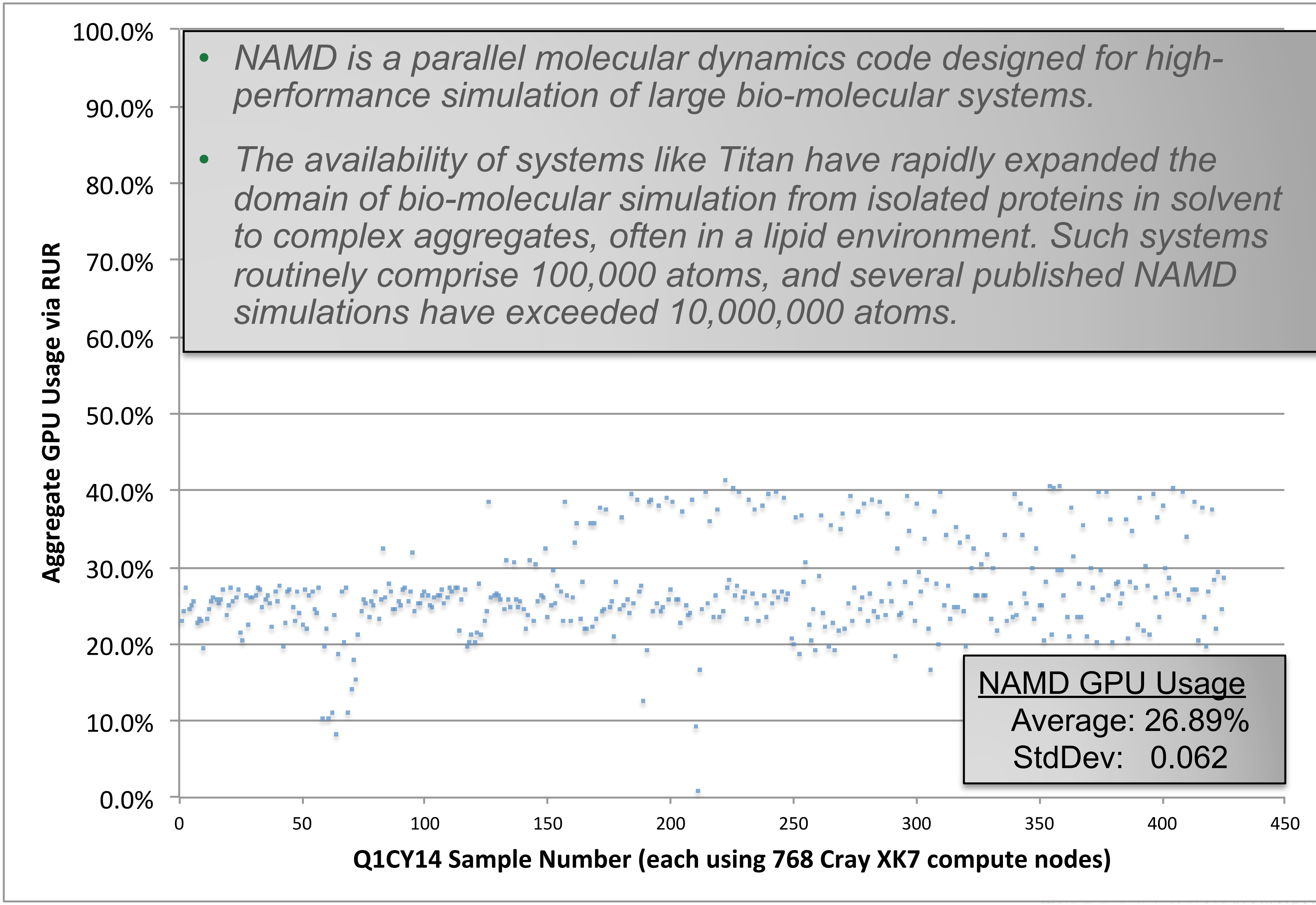
This Series: A sample of all Mixed Mode (OpenMP + MPI) LAMMPS runs in Q1CY14.

Average GPU Usage: 49.28%

GPU Usage by NAMD on OLCF's Cray XK7 Titan

NVML_COMPUTEMODE_EXCLUSIVE_PROCESS

- *NAMD is a parallel molecular dynamics code designed for high-performance simulation of large bio-molecular systems.*
- *The availability of systems like Titan have rapidly expanded the domain of bio-molecular simulation from isolated proteins in solvent to complex aggregates, often in a lipid environment. Such systems routinely comprise 100,000 atoms, and several published NAMD simulations have exceeded 10,000,000 atoms.*



Content

- The OLCF's Cray XK7 Titan
  - Hardware Description
  - Assessing the Operational Impact to Delivered Science
    - Time- and Energy- to Solution. Case Study: WL-LSMS
- The Operational Need to Understand Usage
  - ALTD (the early years)
  - NVIDIA's Role
    - Δ to the Kepler Driver, API, and NVML
  - Cray's Resource Utilization (RUR)
- Examples of NVML_COMPUTEMODE_EXCLUSIVE_PROCESS Measurement
  - Lattice QCD
  - LAMMPS
  - NAMD
- Next Steps…
- INCITE Allocation Program

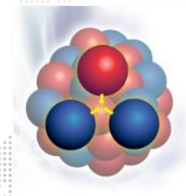
Next Steps

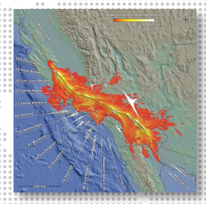
- Clarify mechanisms for correctly understanding and reporting GPU usage as reported for applications using NVML_COMPUTEMODE_DEFAULT.
  - Driver Update? RUR Update?
- Ensure that Memory Usage information reported by NVIDIA Driver is accurate/trustworthy.
  - Driver update?
- Further extend reporting capabilities within the Resource Allocation and Tracking System (RATS) to simplify queries

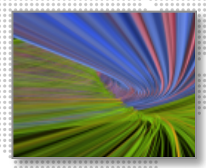
Innovative and Novel Computational Impact on Theory and Experiment

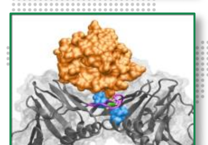
INCITE is an annual, peer-review allocation program that provides unprecedented computational and data science resources

- 5.8 billion core-hours awarded for 2014 on the 27-petaflop Cray XK7 "Titan" and the 10-petaflop IBM BG/Q "Mira"
- Average award: 78 million core-hours on Titan and 88 million core-hours on Mira in 2014
- INCITE is open to any science domain
- INCITE seeks computationally intensive, large-scale research campaigns









Call for Proposals

The INCITE program seeks proposals for high-impact science and technology research challenges that require the power of the leadership-class systems. Allocations will be for calendar year 2015.

Call is Open:
April 16 – June 27, 2014

www.doeleadershipcomputing.org

Contact information
Julia C. White, INCITE Manager
whitejc@DOEleadershipcomputing.org

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY


Questions?