Adaptive COMPUTING

Intelligent Workload Management Software for HPC and Cloud Environments

Accelerate Insights with Topology, High Throughput and Power Advancements

Michael A. Jackson, President
Wil Wellington, EMEA Professional Services

May 2014

Adaptive/Cray Example Joint Customers

Cray Implementations with Over 20,000 Nodes, Topology Aware, Dual Domain, Workload-Aware Power Management

- **NCSA** **(Blue Waters)**
- **Oak Ridge**
- **LANL**
- **Sandia**
- **NOAA**
- **HLRN**
- **UTK**
- **HLRS**
- **ExxonMobil**
- **Univ of Chicago**
- **Laval Univ**
- **Penn State**
- **KTH**
- **ARSC (U of Alaska)**
- **Univ of Bergen**
- **NERSC**
- **Indiana Univ**
- **Colorado State**
- **Tokyo Inst of Tech**
- **Penn State**
- **Texas A&M**

Adaptive Computing Highlights

- **Innovating world-class HPC solutions for over 12 years**
  - Pioneers of HPC schedulers, grid, power management, HPC-Cloud, optimization, scale, dynamic provisioning, Big Workflow and more
  - 50+ patents issued or pending
    - Important for customers concerned about Indemnification risks
  - Backed by top-tier investors

intel Capital

EPIC VENTURES

TUDOR Growth Equity

- **Many customers in the Top 100, including #2 Titan**
  - Largest provider of HPC workload management software to HPC sites*
  - Long history of running the most powerful systems in the world
  - Global partnership with Cray since 2007 – reselling Moab for 7 years

TOP500® SUPERCOMPUTER SITES

THE GREEN 500™

Gartner 2013 Cool Vendor

RED HERRING WINNER 100 N. AMERICA

IDC Analyze the Future

*According to the IDC 2013 HPC End-user Study of System Software and Middleware in Technical Computing

Accelerating Insights with Moab

- **Topology Aware Scheduling**
  - Improve application performance by 2X
  - Based on communication intensity of jobs
- **High Throughput Scheduling**
  - Over 150X more job starts per second
- **Power Savings**
  - Up to 20% Power Savings
  - Reduce carbon usage with less than 5% performance impact
- **30X faster command response on large systems**

- **Better Cray ROI**
  - Faster job launching,
  - faster processing of network-intense workloads,
  - better overall performance means more insights accomplished on the same hardware investment

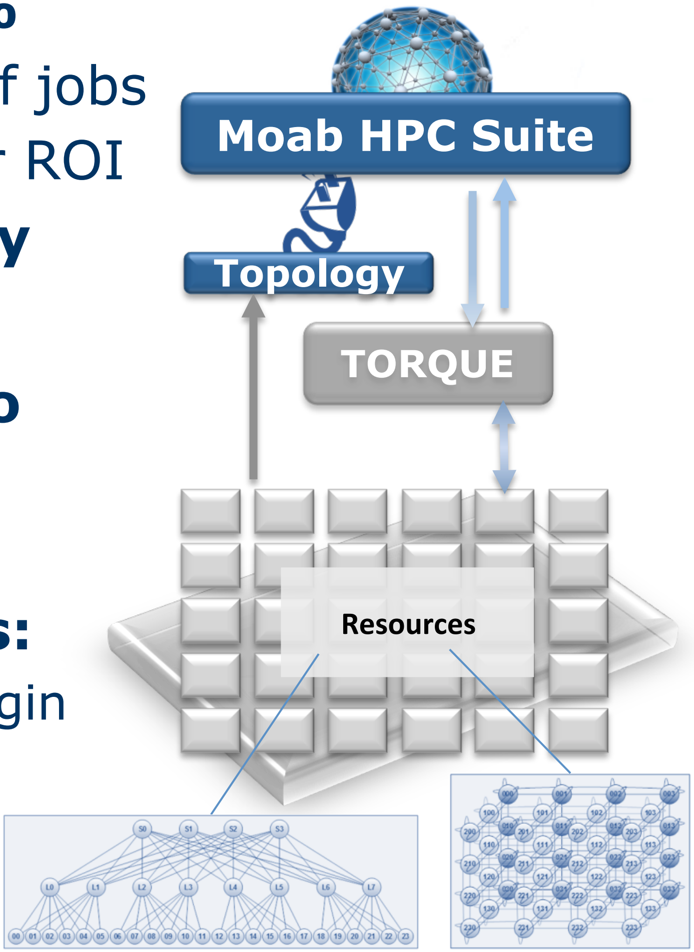
Intelligent Workload Management Software for HPC and Cloud Environments

Topology-aware Scheduling

Faster Processing Due to
Faster Communications

Moab HPC Suite is Optimized for Cray: Topology-based Scheduling Capability

- **Speed job processing up to 200%**
  - Depending on network intensity of jobs
  - Run more jobs per month - better ROI
- **Maintain Job run time consistency with less than 5% variance**
- **Schedule jobs on nodes closest to each other; closer = faster**
- **Topology node allocation plugin capability for different topologies:**
  - Cray ALPS Inventory Topology Plugin currently available
  - Additional Cray-specific plugin
    - 3D Torus
    - Others in development



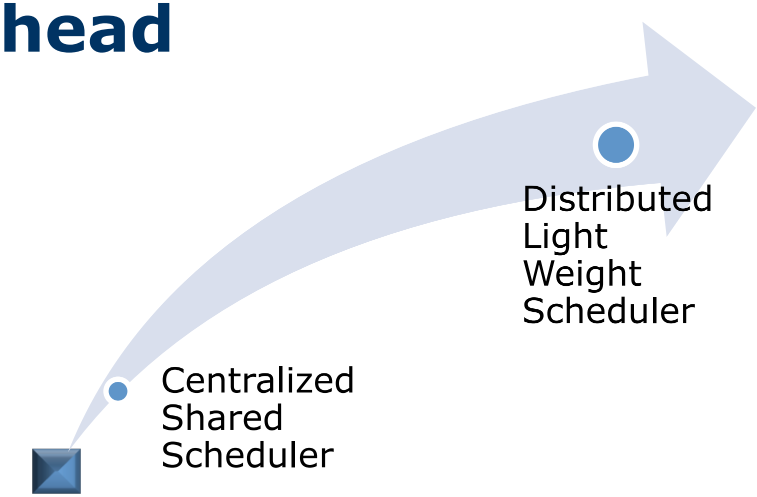
Moab Task Manager

High Throughput Computing
(For TORQUE, Slurm, etc.)

Moab Task Manager (MTM)

High Throughput Computing

- **Distributed lightweight scheduler**
- **Allows 1000’s of job launches per second**
- **Simplifies and offloads global scheduler**
- **No ‘per task’ policy overhead**


Distributed Light Weight Scheduler
Centralized Shared Scheduler

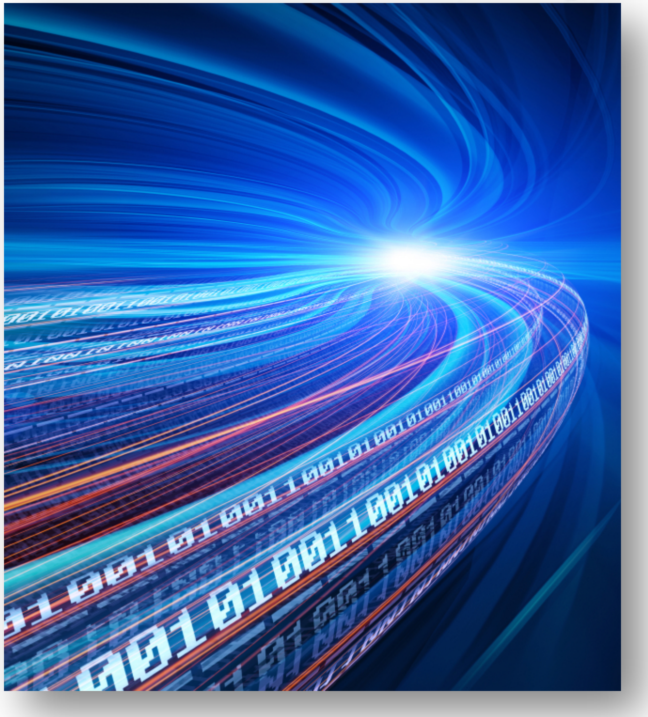
HTC Architecture: 160,000 Jobs Launched / Second*

HPC Architecture: 10-100 Jobs Launched / Second*

*Assuming 1000 Nodes with 16 cores/node

How Does Moab Task Manager Work?

- **Ultra high-speed message queue**
- **Different approach to scheduling**
  - MTM is a transiently invoked sub cluster
  - Combines small, alike jobs to a session
  - Creates policies for the group of jobs
  - Schedules it as one job
  - Incurs scheduling overhead only once, not once per individual small job



- **Limitations**
  - Bounded by processor speed & job size
  - Job I/O requirements may limit speed
  - MTM sacrifices some granularity in management
    - The batch is the unit of management and reporting
    - i.e. individual tasks in a large batch cannot be cancelled or pre-empted in isolation

High Throughput Problem – solved by MTM

Example:

10 Million Jobs on 100 Node Cluster (16 cores/node)

- **HPC scheduler**, at 100 Jobs per second launch rate = **27 hours**
- **Moab Task Manager**, at 10 “tasks”/second/core launch rate
  = **0.17 hours** (Over 150 times faster)

Lab Test Results:
http://www.adaptivecomputing.com/blog-hpc/announcing-early-availability-moab-task-manager/

- 10 Million Jobs on **20 Node Cluster** in 0.21 hours (~13800 tasks / sec)

MTM insertion mode for dynamic workflows

MTM internal workflow

- MTM session submission to Moab:

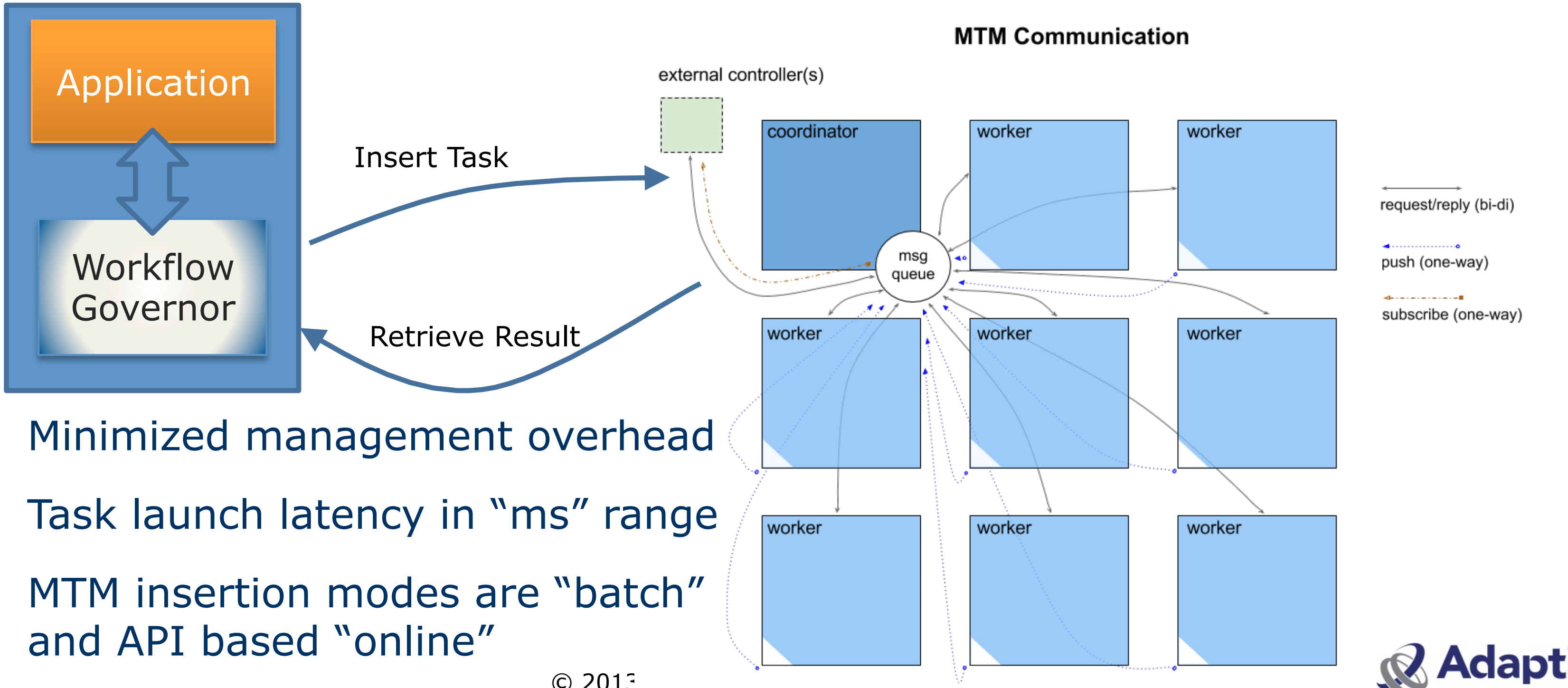
```
msub –l nodes=9 mtm -i tasklist
```

- MTM coordinator launch by Torque:

```
nitro –i --exechosts hostlist tasklist
```

- Task Insertion in to existing session:

```
msub –i <mtm-ID> new-tasklist
```



- Minimized management overhead
- Task launch latency in “ms” range
- MTM insertion modes are “batch” and API based “online”

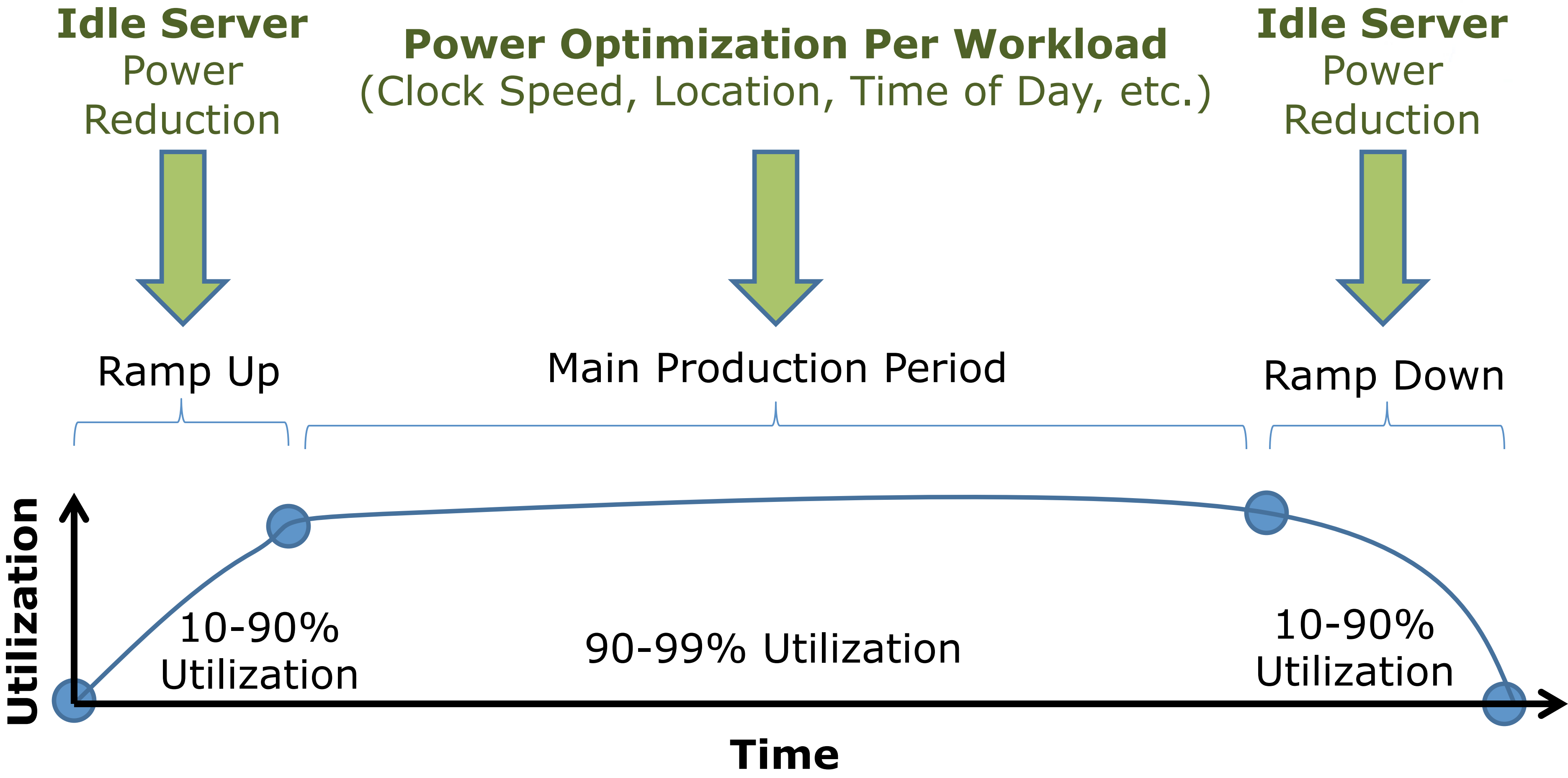
Green Computing

(Includes roadmap features for upcoming June release)

Green Computing – Why

- **Save Power**
  - Limits to Availability
  - Reduce Carbon Emissions
  - Meet Regulations / Goals
- **Save Money**
  - Less Power – Up to 20%
  - Cheaper Power
- **Avoid Overloads**
  - To Grid or Cluster due to Lim

Green Computing – What / When



Idle Server Power Reduction

Power Optimization Per Workload (Clock Speed, Location, Time of Day, etc.)

Idle Server Power Reduction

Ramp Up

Main Production Period

Ramp Down

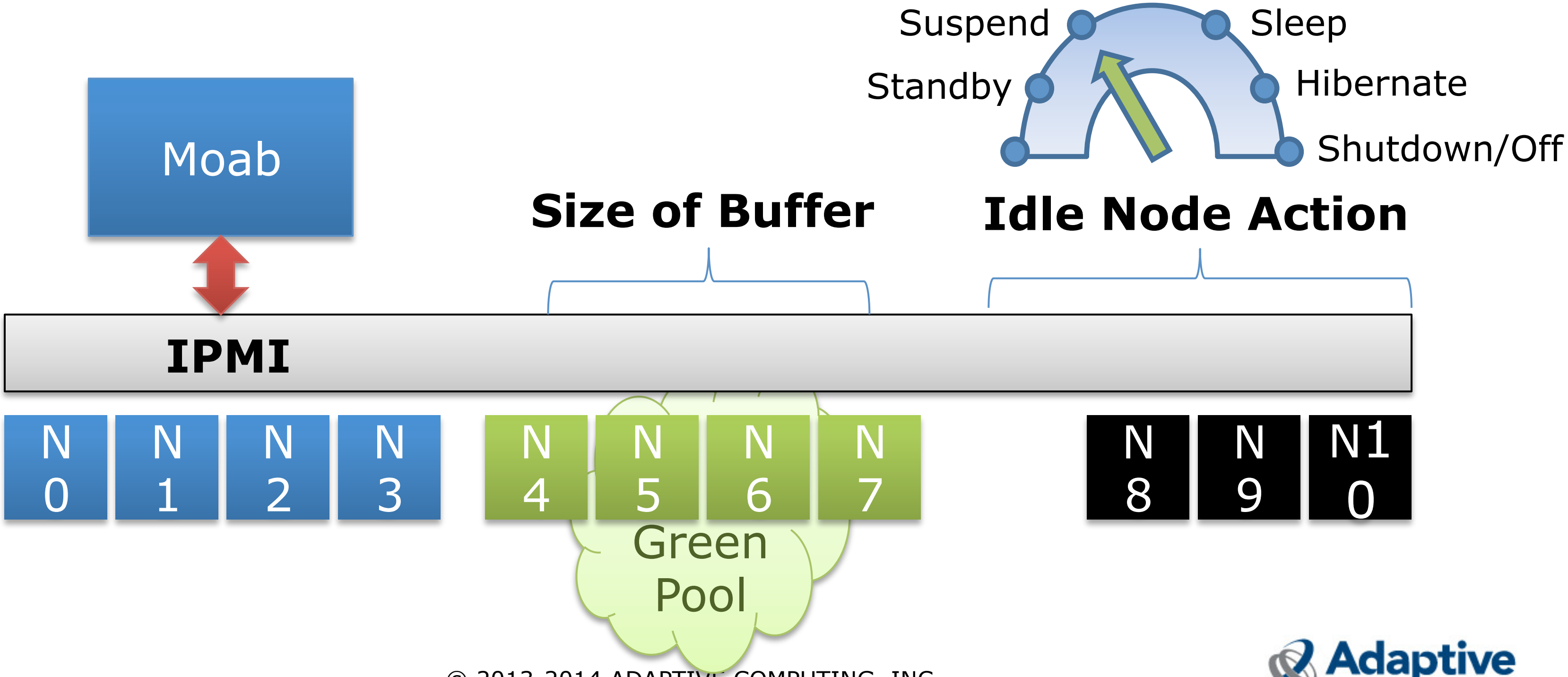
10-90% Utilization

90-99% Utilization

10-90% Utilization

Idle Server Power Reduction

- **Save energy costs reducing power on idle nodes**
- **Maintain response time with Green Pool Buffer Policy**
- **Reference scripts provided (OpenIPMI)**

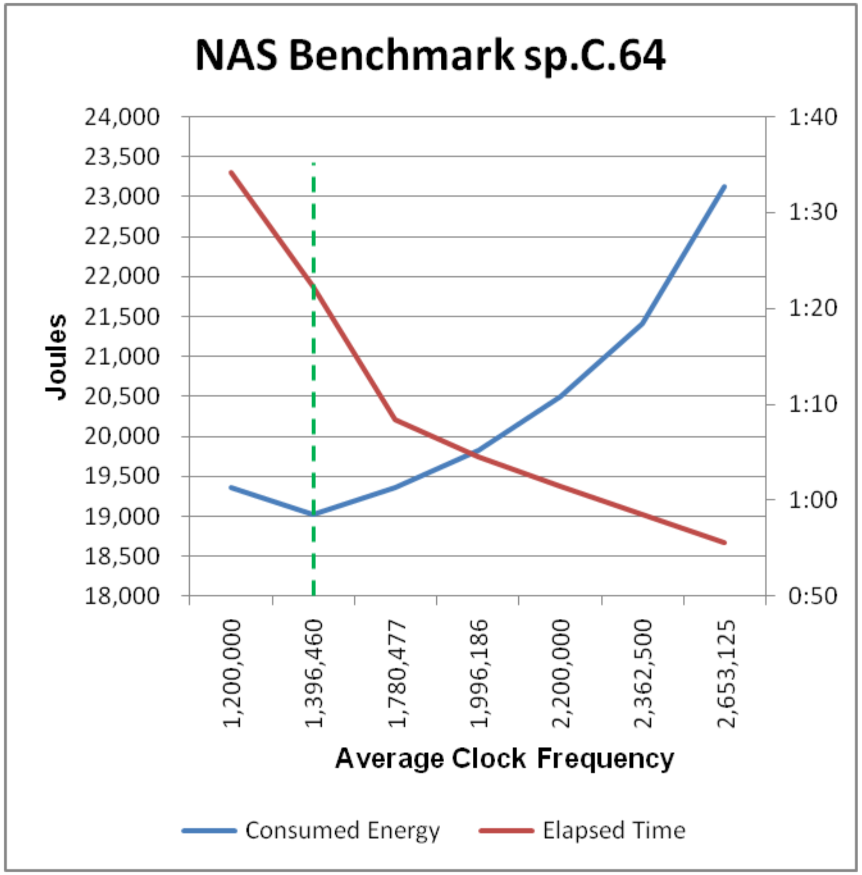


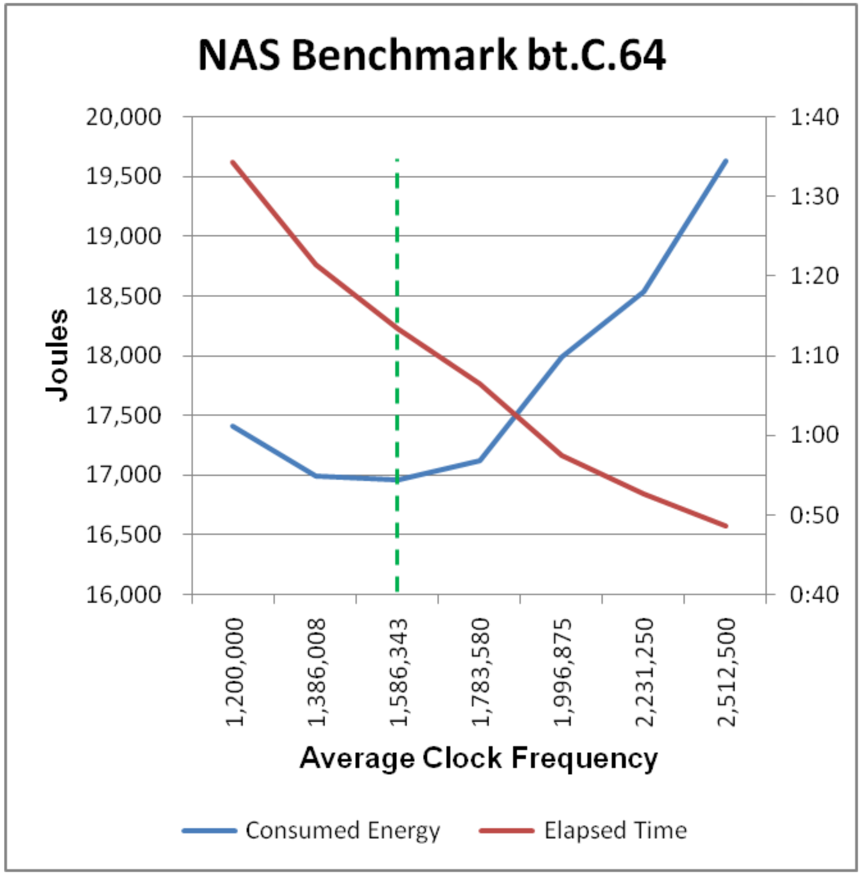
Power Optimization Per Workload

- **When utilization is high, focus on power optimization per workload**
  - Analyze:
    - Completion Time Goals
    - Workload Energy/Runtime Profiles
    - Energy Costs
  - Optimize:
    - Energy Consumption vs. Target Job Run Time

Energy/Runtime Profiles

- **Minimizing energy consumption requires application-specific optimal clock frequency**





CPU Clock Frequency Control

- **New cpuclock= job submission option**
  - Absolute Clock Frequency Number
    - Example: `cpuclock=2200` or `cpuclock=1800mhz`
  - Linux Power Governor Policy
    - Example: `cpuclock=conservative`
  - Relative P-state Number *(not available for XC/XK/XE)*
    - Values 0-15
      - 0="turbo" frequency
      - 15=slowest frequency
    - Example: `cpuclock=0` or `cpuclock=P2`
- **Can set in job templates**

Green Computing Thought Leadership and Indemnification

Adaptive Computing has Thought Leadership and Intellectual Property in Green Computing

- **Analyze**: Workload (Current and Future), Resource State, Energy Consumption, Temperature, Energy Costs, Aggregate Energy Use, Time of Day, Location, etc.
- **Modify**: Power State, Clock Speed, Placement, etc.
- **Patents**:

8,276,008 B2	8,245,059 B2
8,549,333 B2	8,271,807 B2
8,271,813 B2	

INTELLECTUAL PROPERTY

Patented

ALL RIGHTS RESERVED

- **Indemnification**:
  Adaptive Computing indemnifies users/vendors on Moab Workload Management uses for green computing.

New Capabilities in the Next Release

Scale Large System Responsiveness (Size and Speed)

- **3.5X to 4X faster**
  - Moab scheduling speed on very large systems
  - Better multi-threading of non-scheduling services
- **30X+ faster command responsiveness**
  - (showq, mdiag, showres, showstart, showbf, checkjob, checknode, showstats)
  - Low Latency Command Initiative
- **2X+ improvement in TORQUE job communication handling**
  - more jobs
  - more job starts
  - more job exits

Grid Job Scheduling

Job Information

- 30 minute Job Walltime Estimate
- 2.5 GB input file(s) size
- 1.5 GB output file(s) size

Job Wait Time (from now)

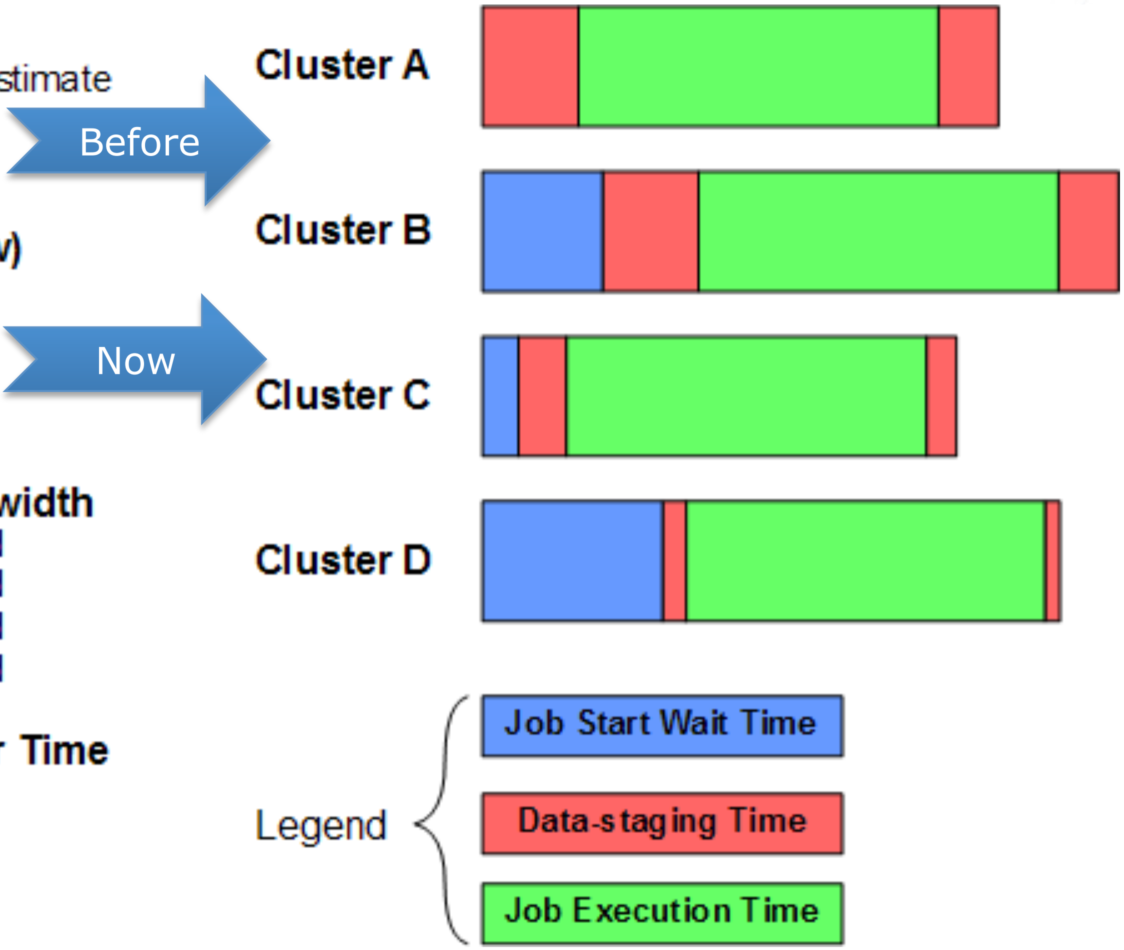
- Cluster A – 0 seconds
- Cluster B – 600 seconds
- Cluster C – 180 seconds
- Cluster D – 900 seconds

Available Network Bandwidth

- Cluster A – 5 MB/second
- Cluster B – 5 MB/second
- Cluster C – 10 MB/second
- Cluster D – 20 MB/second

Calculated Data Transfer Time

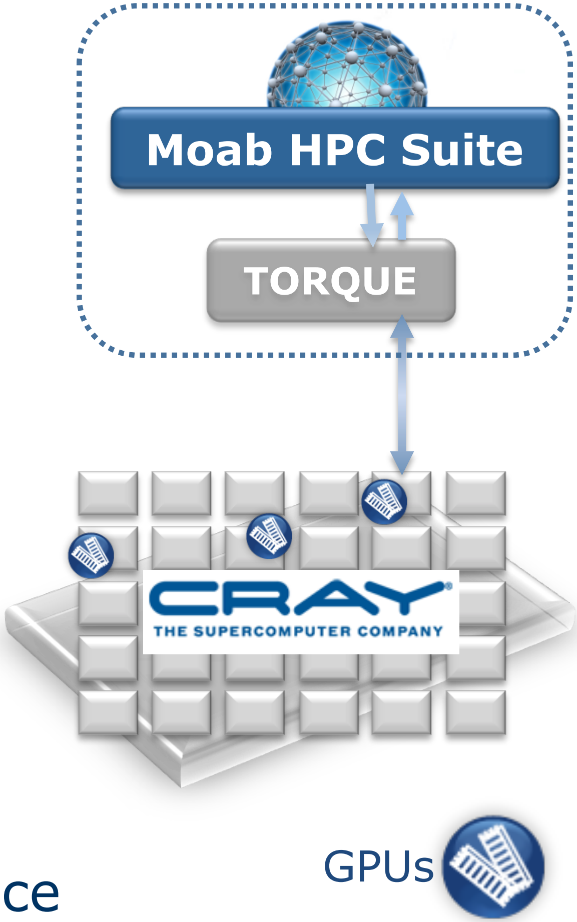
- Cluster A – 800 seconds
- Cluster B – 800 seconds
- Cluster C – 400 seconds
- Cluster D – 200 seconds



Moab HPC Suite is Optimized for Cray: Faster, More Reliable Scheduling for Cray

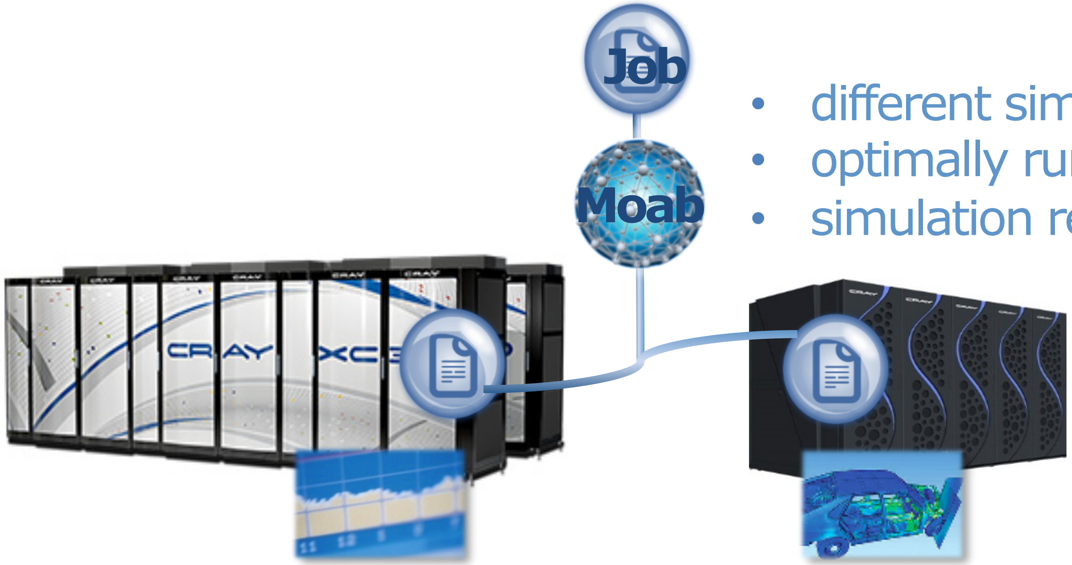
Streamlined Moab HPC Suite and Cray ALPS architecture via external server

- Increased demands on the scheduler don’t impact other SDB processes
- “Beefed up” external Moab server for faster scheduling
- Moab and TORQUE can be run in high availability mode for robustness
- Submit/query jobs during Cray maintenance/downtime
- Better ALPS reservation cleanup
- Auto-detection of Cray nodes and accelerators
- Faster deployment with simpler interface



Moab HPC Suite is Optimized for Cray: Dual Domain Job Scheduling for Cray

- **Speed job submission and results**
- **Schedule single job, runs simultaneously across Cray HPC and Cray Cluster or non-Cray compute nodes**
  - no wasted duplicate job submission
  - no waiting to submit dependent job to second domain



- different simulation/analysis
- optimally run on different resources
- simulation results can still interact

Accelerating Insights with Moab

- **Topology Aware Scheduling**
  - Improve application performance by 2X
  - Based on communication intensity of jobs
- **High Throughput Scheduling**
  - Over 150X more job starts per second
- **Power Savings**
  - Up to 20% Power Savings
  - Reduce carbon usage with less than 5% performance impact
- **30X faster command response on large systems**

- **Better Cray ROI**
  - Faster job launching,
  - faster processing of network-intense workloads,
  - better overall performance means more insights accomplished on the same hardware investment

Questions?