E2 HPC Working Group

SC16 Salt Lake City, Utah | hpc matters.

Sustainably Supporting Science
through committed community action

Workload Management Panel

SC16 Energy Efficient HPC Working Group Workshop
November 13, 2016

E2 HPC Working Group

SC16 Salt Lake City, Utah | hpc matters.

Sustainably Supporting Science
through committed community action

Moderator: Steven Martin, Cray Inc.

Question:

Please describe an energy-efficiency or power management use-case that your customers are demanding a solution for?

E2 HPC Working Group

SC16 Salt Lake City, Utah | hpc matters.

Sustainably Supporting Science
through committed community action

Agenda:

Each panel member will have 5 min for prepared slides

The rest of the hour will be Q/A with the audience

Workload Management Panel

- Khalid Ahmed: IBM, IBM Spectrum LSF
- Gary Brown: Adaptive Computing, Moab / Torque
- Jonathan Shelley: Altair, PBSPro
- Ian Lumb: Univa, Grid Engine
- Tim Wickberg: SchedMD, Slurm

Workload Management Panel

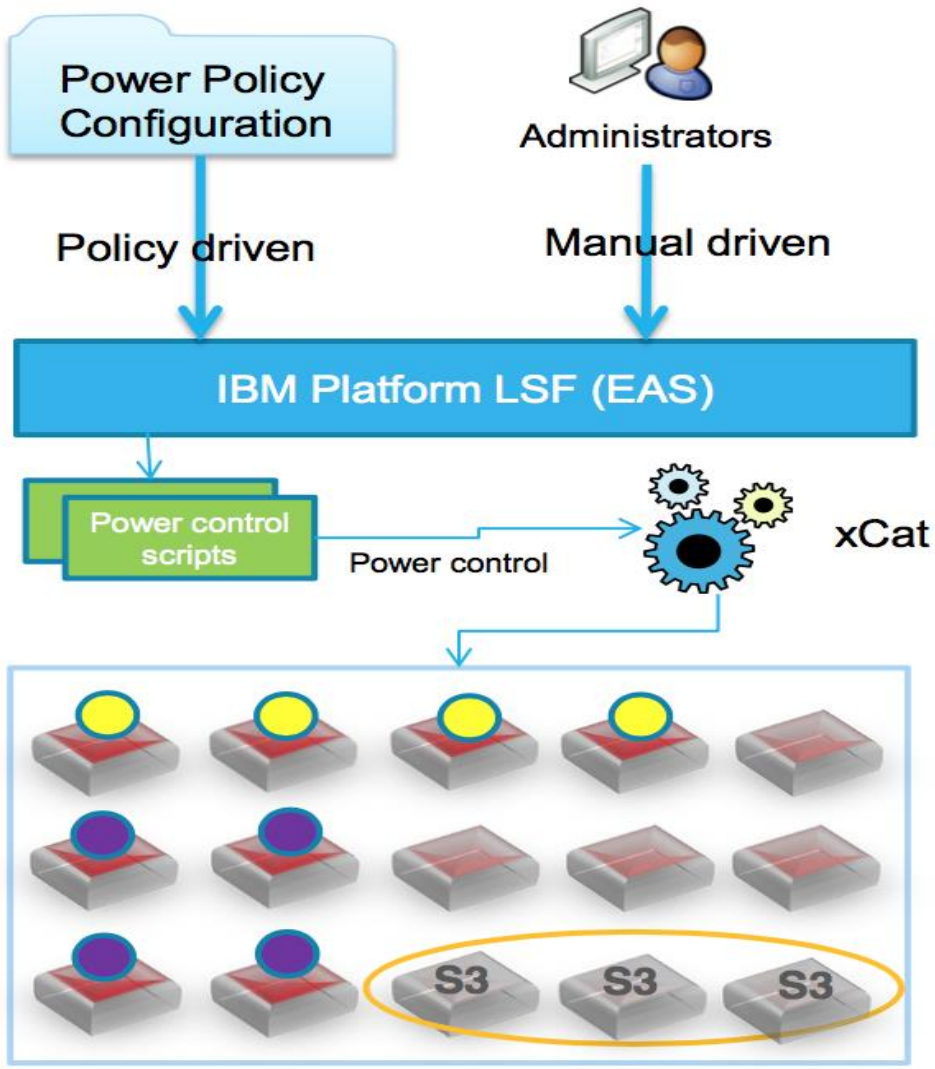
Khalid Ahmed: IBM, IBM Spectrum LSF

Working Group

Energy Efficiency Requirements

- Reduce power consumption on idle nodes
- Optimize power consumption of active nodes for running workloads
- Provide energy consumption accounting per job
- Power optimization across multiple HPC clusters
- Support heterogeneous hardware configurations
  - Nodes with and without accelerators in the same cluster
  - Different hardware, e.g. x86_64, POWER8, ARM, typically in different clusters, which may be connected via LSF MultiCluster
  - Nodes are not equal: seemingly same nodes may behave differently with regard to power consumption and performance.

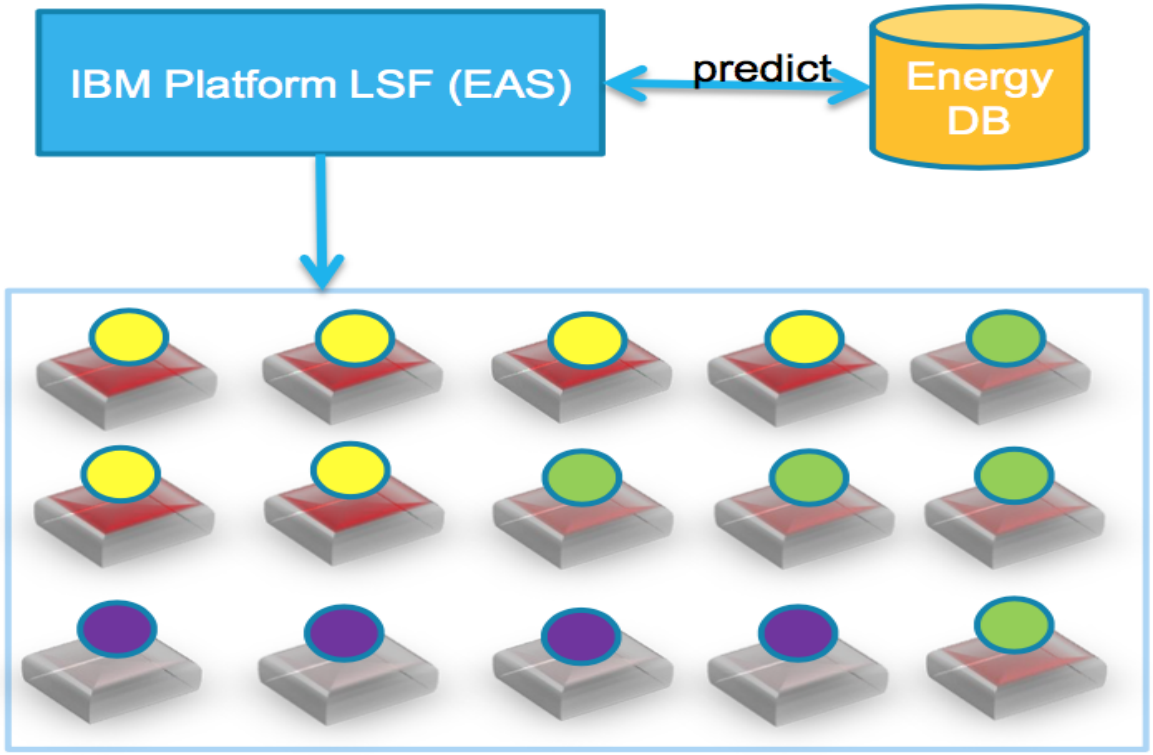
How LSF supports Host Power Management?



- **Manual management tools**
  - badmin hpower suspend/resume host
  - badmin hist
- **Policy Driven Power Saving**
  - Policy windows (i.e., 10:00 PM – 7:00 AM)
  - Node will be put in power saved (S3) if it is idle for a configurable period of time.
- **Power Saving Aware Scheduling**
  - Schedule jobs to use idle nodes first (Power saved nodes as last resort)
  - Aware of job request and wake up nodes precisely on demand
  - Safe period before running job on resumed nodes
- **Others**
  - customizable power control scripts
    - OOB xcat support
  - Power events tracking

How LSF Automatically Select Optimal CPU frequency

- Step I: Learning
  - LSF evaluates the power profile of all nodes
  - calculates coefficents factors
  - save them in the energy database



- Step II: Set Default Frequency
  - System administrator defines cluster default cpu frequency (nominal or lower frequency)
- Step III: Tag the job first time
  - User submits the application with a tag
  - runs the job under default frequency
  - LSF collects energy, power, runtime, hardware counters
  - Generates predication result and saves in database
- Step IV: Use predication
  - User re-submits the same application with the same tag and specifies energy policy
  - LSF selects the optimal cpu frequency for application based on predication result and policy setting.
  - Run the application under selected frequency

More to come…

The paper to be presented at the E2SC Workshop on Monday, Nov. 14.

V.Elisseev, J.Baker, N.Morgan, L.Brochard, T.Hewitt "Energy Aware Scheduling Study on BlueWonder"
To appear in Proc. Of 4th Workshop on Energy-efficient Supercomputing (E2SC), held in conjunction with the IEEE/ACM Conference on High Performance Computing, Networking, Storage, and Analysis (SC'16), Salt Lake City, UT, Nov. 2016.

Abstract:
Power consumption of the world's leading supercomputers is of the order of tens of MegaWatts (MW). Therefore, energy efficiency and power management of High Performance Computing (HPC) systems are among the main goals of the HPC community. This paper presents our study of managing energy consumption of supercomputers with the use of the energy aware workload management software IBM Platform Load Sharing Facility (LSF). We analyze energy consumption and workloads of the IBM NextScale Cluster, BlueWonder, located at the Daresbury Laboratory, STFC, UK. We describe power management algorithms implemented as Energy Aware Scheduling (EAS) policies in the IBM Platform LSF software. We show the effect of the power management policies on supercomputer efficiency and power consumption using experimental as well as simulated data from scientific workloads on the BlueWonder supercomputer. We observed energy saving of up to 12% from EAS policies.

Workload Management Panel

Gary Brown: Adaptive Computing, Moab / Torque

Working Group

“Power Band” and “Ramp Rate” Use Case

Question: What do you have to do if your HPC system’s power load can materially affect your electrical power provider’s generation, transmission, and/or distribution equipment?

Gary D. Brown, Adaptive Computing

Product Manager for Moab Workload Manager job scheduler and open-source TORQUE Resource Manager

Adaptive
COMPUTING

Why “Power Band” and “Ramp Rate” Control?

- Trinity HPC system at Los Alamos National Laboratory (LANL)
  - Significant percentage of LANL’s total power requirement
- DoE LANL located in Los Alamos County, New Mexico
  - Los Alamos County Dept. of Public Utilities is an electrical power provider
- LAC and LANL entered into an “Energy Coordination Agreement” (ECA)
  - Coordinate use of LAC and LANL power generation equipment
  - Coordinate power purchase by LANL from LAC and the “open market”
- LANL uses ~80% of ECA power

Power Sources (NREL 2015 LAC-DPU study)

- Los Alamos County, New Mexico (LAC)

LAC Source	Capacity (min-max in 2009-2014)
Hydro	1-24 MW
Coal-fired	46 MW
Solar	0-1 MW
Total	**47-71 MW**

- Dept. of Energy's Los Alamos National Laboratory (LANL)

LANL Source	Capacity
Hydro (WAPA)	11 MW
Gas-fired*	32 MW (* local backup/emergency)
Total	**11-43 MW**
Must purchase*	17-49 MW (* ECA + "open market")

Los Alamos County / DoE LANL Power Use

- Los Alamos County Power *(2000-2014 average)*
  - Summer Daily Load Range 12-19 MW
  - Winter Daily Load Range 13-18 MW
  - Spring/Fall Daily Load Range 11-14 MW
- DoE LANL Power *(typical but can be variable!)*
  - Summer Daily Load Range 54-60 MW
  - Winter Daily Load Range 56-59 MW
  - Spring/Fall Daily Load Range 38-44 MW
- Total ECA (LAC + LANL) Power
  - Summer Daily Load Range 66-79 MW
  - Winter Daily Load Range 69-77 MW
  - Spring/Fall Daily Load Range 49-58 MW

Trinity and LANL Power Contract

- Trinity's power use is 8-10 MW
  - 13-26% of possible LANL load ranges
- Power contract requires advance notice of anticipated power load 24-72 hours in advance
  - Permits power use within "power band" of 2 MW centered on power load
  - Imposes power ramp-rate limit of 1 MW per minute within power band
- Power contract penalties
  - Exceed power band maximum limit
    - 2x normal cost if provider has available power
    - Up to 300x normal cost if must purchase power on "spot" market
  - Drop below power band minimum limit
  - Exceed ramp-rate limit

Power Band and Ramp-Rate Management

- Working with LANL and Sandia National Laboratory
  - Power-aware job scheduling and compute node power management policies
  - Power-Band management policies
  - Ramp-rate management policies
  - Job submission, job script, job template, and QoS power-management options
- Controlling power use with Cray's "capmc" utility
  - Control idle compute node "global node power" state (G-state)
    - Power On / Shut Down / Reboot
  - Control idle / active compute node "CPU" state (C-state)
    - Active / Halt / Sleep / Deep Sleep
  - Control idle / active compute node "Performance" state (P-state)
    - Clock Frequency / Power Cap
- Anticipate working with "PowerAPI" utilities

Workload Management Panel

Jonathan Shelley: Altair, PBSPro

Working Group

Introduction

- Jon Shelley
  - Director PBS Professional
- Background
  - 3.5 years at Altair
  - 6 years managing HPC resources at the Idaho National Lab
- Hobbies
  - Biking
  - Walking the dogs
  - Family

- Altair
  - Founded in 1985
  - 48 offices across 20 countries
  - 2500+ employees world wide
  - Acquired PBS Professional in 2003
  - Started delivering power features in 2008

Altair

Energy Efficient Requests

- Power cap the node/system when capacity is reached
  - Implemented on Cray XC, SGI ICE, and other partner
- Ramp down the system when cooling is inadequate
- Job and node level power ramp up and down rate.

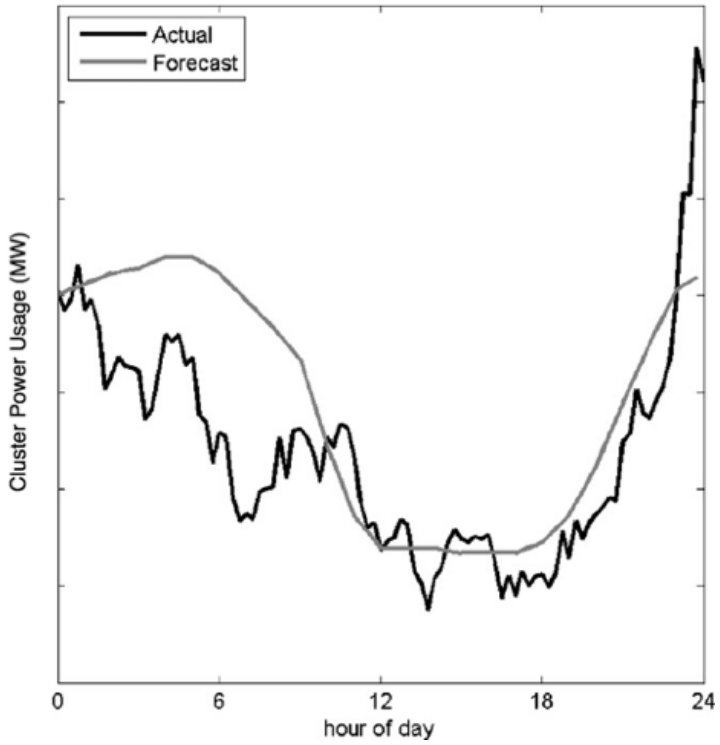
"Energy costs are the fastest-rising cost element in the data center."

- Gartner

Altair

Energy Efficient Requests Cont…

- Per job energy accounting
  - Implemented on Cray XC, SGI ICE, and other partner
- Power profiles/templates
  - Implemented on Cray XC, SGI ICE, and other partner
- System and job level profiling and forecasting.
  - e.g. prediction of hourly power averages >24 hours in advance with a 1 MW tolerance.
  - http://www.lanl.gov/projects/apex/_assets/docs/APEX2020_draft_tech_specs_v4.1.pdf



Altair

Workload Management Panel

Ian Lumb: Univa, Grid Engine

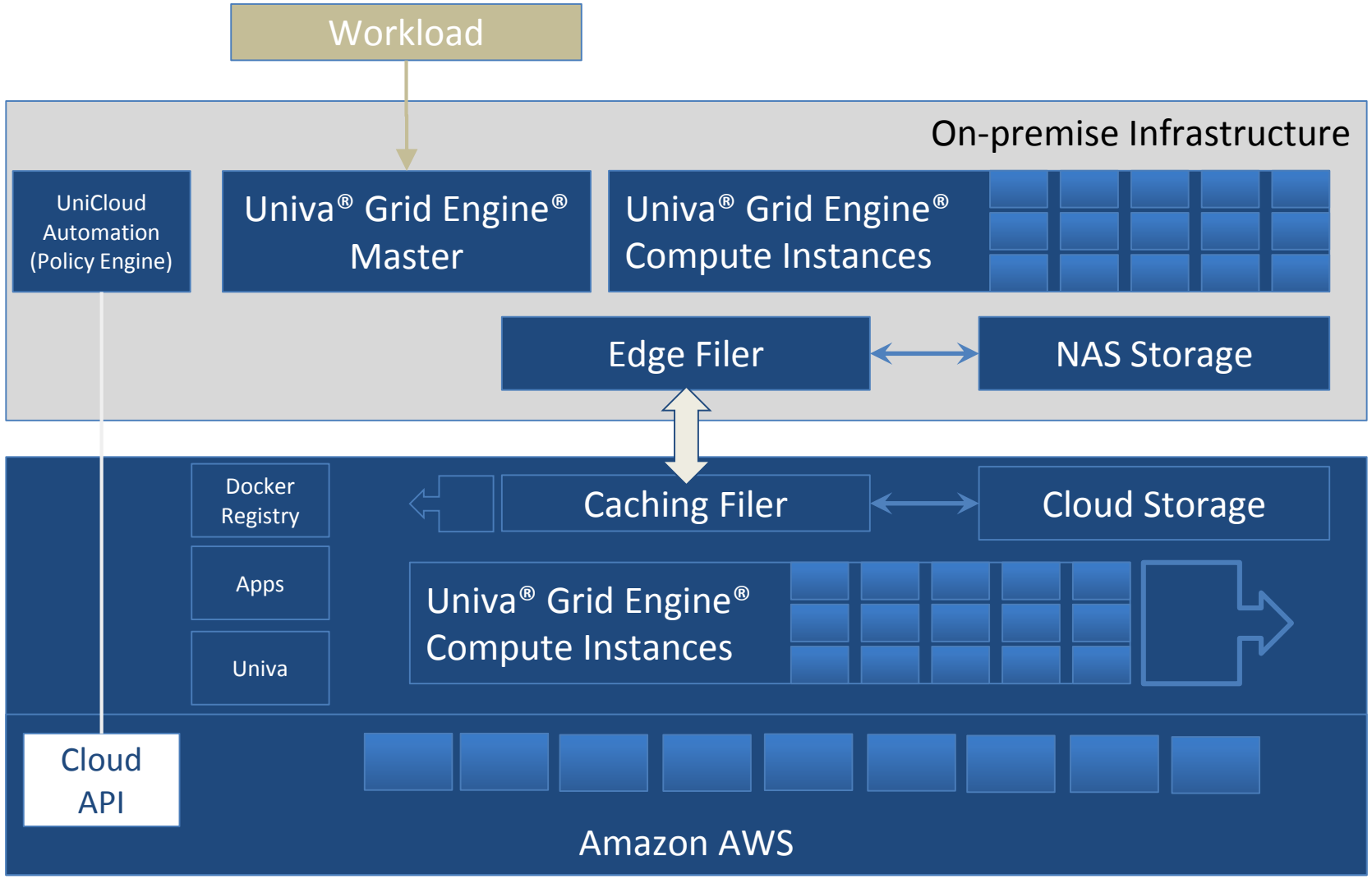
Working Group

USE-CASE CONSIDERATIONS

- Capacity and capability computing can result in radically different use cases
  - Univa has more experience with capacity computing use cases aimed at maximizing/optimizing for (peak) throughput in clusters and clouds
    - Includes dynamically instantiated resources
      - … in public and private clouds
      - … that ***pull*** - rather than receive *pushed* workload
  - Efficacy of capability computing even more critically dependent upon *a priori* power-consumption estimates – only achievable from benchmarking studies ???
    - Scheduling policies (e.g., Advance Reservation, Preemption) may need to be refactored to incorporate energy efficiency and/or power-management considerations
- Metrics need to be *consistently* exposed for all resources under WLM
  - … and some need to be dynamically adjustable by the WLM
    - WLMs *may* need to work in tandem with cluster managers and even HPC tools
- Refactoring aimed at systematically introducing micro services unrealistic for HPC
  - However, 'pragmatic meso services' may introduce energy efficiencies …
    - Containers are more energy efficient than VMs … which are more EE than physical machines …

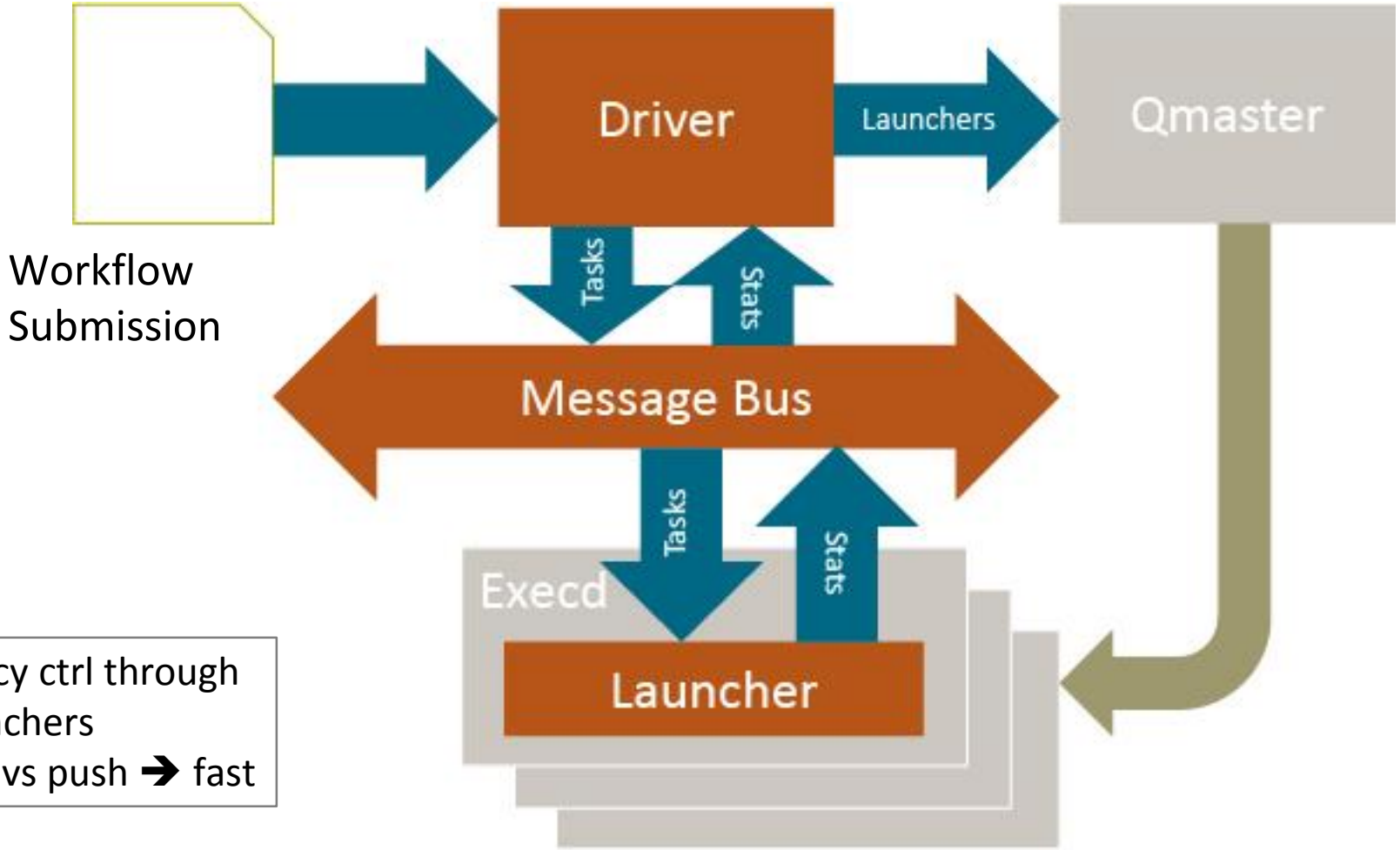
Copyright © 2016 Univa Corporation, All Rights Reserved.

Cloud Bursting: Peak Throughput Supplanted by EE



UNIVA

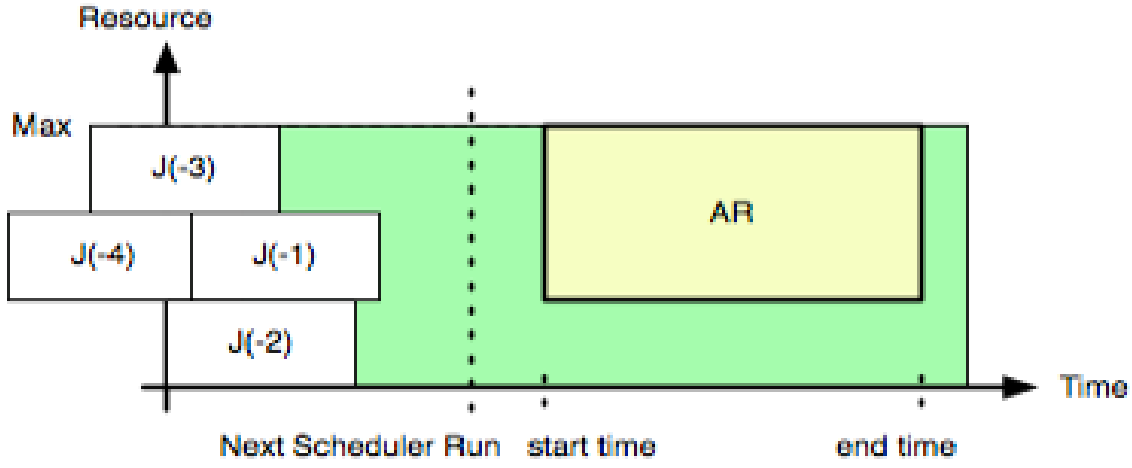
Pulled vs. Pushed Workload



- Policy ctrl through launchers
- Pull vs push ➔ fast

Copyright © 2014 Univa Corporation. All Rights Reserved.

ADVANCE RESERVATION



- Advance Reservation Objects
  - Start Time (Defined during creation)
  - Duration / End Time (Defined during creation)
  - Reserve a defined set of resources
- After creation of an AR Jobs can be submitted into an AR

Copyright © 2016 Univa Corporation, All Rights Reserved.

Workload Management Panel

Tim Wickberg: SchedMD, Slurm

Working Group

SC16 EEHPC WLC Panel

Tim Wickberg
tim@schedmd.com

Copyright 2016 SchedMD LLC
http://www.schedmd.com

Power management use cases

- Power capping - set max power consumption of the system
- React to cooling system performance,
- Potential to react to energy market fluctuations
  - Shed load during peak grid demand, can be a huge cost savings
  - Peaking plants are “dirty”, keep them offline as an environmental win
- Slurm has a reactive power capping system in place today
  - But it could be significantly better

Copyright 2016 SchedMD LLC
http://www.schedmd.com

Improvements to power capping

- Faster polling and more accurate data collection
- Predictive load scheduling
  - Past behavior is not factored into future scheduling decisions
  - Correct mix of jobs could avoid power capping entirely
- Implement a “node shedding” factor
  - When is it better to have fewer online nodes at full speed, rather than all nodes with a lower performance limit?

Copyright 2016 SchedMD LLC
http://www.schedmd.com

E2 HPC Working Group

SC16
Salt Lake City, Utah | hpc matters.