CRAY

# HPM2016:

## HPC Power Management: Knowledge Discovery

**Panel Discussion:** Steven J. Martin (stevem@cray.com)

# Legal Disclaimer

*Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.*

*Cray Inc. may make changes to specifications and product descriptions at any time, without notice.*

*All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.*

*Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.*

*Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.*

*Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.*

*The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, and URIKA. The following are trademarks of Cray Inc.: APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, REVEAL,THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.*

# KAUST Power Capping Summary

- **Shaheen2: 36 cabinet Cray XC40, #10 top500 June2016**
- **Constrained by site power/cooling availability**
  - During acceptance: 2.9 MW limit
  - After acceptance: 2.3 MW limit
- **Two power capping approaches:**
  - Two static queues
  - Dynamic capping with Slurm
- **System and application power profiling used heavily to:**
  - Tune Slurm dynamic power capping
  - Tune application and identify performance problems early in runs
  - Monitor cabinet and system level power usage

# KAUST: Shaheen2, 36 Cabinet Cray XC40

- **6174 dual socket Haswell nodes, 32 cores/node**
  - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect

- **Theoretical peak performance: 7.2 PF**

- **Shaheen2 on Top 500: www.top500.org/system/178515**

| List | Rank | Cores | Rmax | Rpeak | Power (KW) |
|---|---|---|---|---|---|
| 06/2016 | 10 | 196,608 | 5,537.0 | 7,235.2 | 2,834.0 |
| 11/2015 | 9 | 196,608 | 5,537.0 | 7,235.2 | 2,834.0 |
| 06/2015 | 7 | 196,608 | 5,537.0 | 7,235.2 | 2,834.0 |

# KAUST: Constraints

| | During Acceptance | After Acceptance | Peak |
|---|---|---|---|
| Power<br>Cooling | Allocated 2.9 MW<br>- When others systems off/idle | Allocated 2.3 MW | 2.94 MW running<br>- LINPACK + 2 apps across full machine |
| Capping | Two Static Queues<br>- 1805: Nodes Uncapped<br>- 4367: Nodes capped @ 270W | Slurm Dynamic | Disabled |
| Notes: | Data center capacity:<br>- Cooling 2.9 MW<br>- Power ~ 3.2 MW | Systems:<br>- Shaheen2<br>- BG/P 16 racks (~ 500 KW)<br>  - Decommissioned end of 2015<br>- Several other small clusters | |

# KAUST: Two Power Capping Approaches

| | Two static queues: | Dynamic capping with Slurm |
|---|---|---|
| Pros | • Performance reproducibility | • Better utilization and distribution of power across nodes<br>• Reduced time for production runs |
| Cons | • Large scale code cannot run<br>• Lower overall utilization | • High variability of performance<br>  • Up to 2x for compute bound applications when machine is used more than 50% |
| Notes | • 1805: Uncapped nodes<br>• 4367: Nodes capped at 270W<br>• Capped queue up to 2X slower<br>• Users prefer uncapped nodes | • Monitoring used to tune Slurm<br>• Ability to dynamically disable power capping |

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

SHAHEEN
SUPERCOMPUTING LABORATORY

Thanks to Bilel Hadri bilel.hadri@kaust.edu.sa for help pulling together information needed for this presentation.

# Backup Slides

# How the site used power monitoring data?

- **Real-time system power data available**
  - Used by sys-admin, computational scientist, and data centers admins
- **Power profiling of applications, especially the full scale ones**
  - Used when strategizing/optimizing full scale Gordon Bell runs on Shaheen2
- **Detecting issues on applications performance**
  - Known compute intensive code drawing less than 200W per node
    - Found issue in the communication pattern
  - During acceptance runs
    - No need to wait for 40 minutes for a first performance number when the power per cabinet was less than 55KW, while it should operate in the 80s KW

# KAUST: Power and Cooling Constraints

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

- **Data center capacity:**
  - Cooling 2.9 MW
  - Power ~ 3.2 MW
- **Systems:**
  - Shaheen2
  - BG/P 16 racks (~ 500 KW)
    - Decommissioned end of 2015
  - Several other small clusters

# KAUST: Shaheen2 Constraints

- **Shaheen2 during acceptance:**
  - Allocated 2.9 MW
- **Shaheen2 after acceptance:**
  - Operating with 2.3MW power/cooling limit
- **Shaheen2 reached a peak of  2.94MW**
  - LINPACK + 2 other applications
  - Running at full scale across the machine

# Static Queues

- **Two static queues: (using CAPMC)**
  - 1805: Uncapped nodes (allowed to run at full potential)
  - 4367: Nodes capped at 270W
- **Pros:**
  - Performance reproducibility
    - Capped queue is up to 2X slower for some applications
- **Cons:**
  - Large scale code cannot run
  - Lower overall utilization since waiting is longer
    - Users tend to prefer uncapped nodes

# Dynamic capping with SLURM

- **Pros:**
  - Better utilization and distribution of power across the nodes
  - Reduced time for production runs vs static capping at 270 watts
- **Cons:**
  - High variability of performance
    - Up to 2x  for compute bound applications when system load > 50%
- **Notes:**
  - Used Cray monitoring to tune SLURM parameters
    - Improve utilization and distribution of allocated power
  - Ability to dynamic disable power capping on the fly
    - Dedicate the machine up to 75%, Idle the rest
    - Ability to change power limits in case of maintenance or issue with cooling