Power Aware Heterogeneous Node Assembly

Bilge Acun, Alper Buyuktosunoglu, Eun Kyung Lee, Yoonho Park
IBM T. J. Watson Research Center
Outline

1. Motivation
2. Power Variation Analysis
3. Variation Aware Node-Assembly Techniques
4. Evaluation
5. Conclusion
Outline

1. Motivation
2. Power Variation Analysis
3. Variation Aware Node-Assembly Techniques
4. Evaluation
5. Conclusion
Motivation: Heterogenous Fat Compute Nodes

www.top500.org
November, 2018 list

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>System</th>
<th>Cores</th>
<th>Rmax</th>
<th>Rpeak</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/SC/Oak Ridge National Laboratory</td>
<td>Summit - IBM Power System</td>
<td>2,397,824</td>
<td>143,500.0</td>
<td>200,794.9</td>
<td>9,783</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td>AC922, IBM POWER9 22C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.07GHz, NVIDIA Volta GV100,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dual-rail Mellanox EDR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infiniband</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IBM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>DOE/NNSA/LLNL</td>
<td>Sierra - IBM Power System</td>
<td>1,572,480</td>
<td>94,640.0</td>
<td>125,712.0</td>
<td>7,438</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td>S922LC, IBM POWER9 22C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.1GHz, NVIDIA Volta GV100,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dual-rail Mellanox EDR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infiniband</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IBM / NVIDIA / Mellanox</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>National Supercomputing Center in Wuxi</td>
<td>Sunway TaihuLight - Sunway</td>
<td>10,649,600</td>
<td>93,014.6</td>
<td>125,435.9</td>
<td>15,371</td>
</tr>
<tr>
<td></td>
<td>China</td>
<td>MPP, Sunway SW26010 260C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.45GHz, Sunway</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NRCPC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Motivation: Manufacturing Variations in Hardware

- Parametric data of 190 IBM POWER8 chips showing the correlation between quiescent current (Iddq) and PSRO (performance sort/screen ring oscillator).

- Supply voltage distribution fitting a Gaussian distribution.
Motivation: Insufficient Scheduling Methods

• Power aware job scheduling comes with a performance trade-off
  • Contiguous node allocations are used to optimize for network performance
  • Moving the threads can be bad for locality

• Supercomputer job schedulers cannot address within node variations
  • Nodes are allocated exclusively to each application
  • Good and bad chip might end up in the same node
Variation Aware Node-Assembly Methods

Illustration of Type-1 Node Assembly

Illustration of Type-2 Node Assembly

Illustration of Type-3 Node Assembly

Sorted
 Balanced
 App-Aware
Outline

1. Motivation

2. Power Variation Analysis

3. Variation Aware Node-Assembly Techniques

4. Evaluation

5. Conclusion
Static Power Distribution

- We use the open-source AMESTER tool in order to make voltage, power and temperature measurements in IBM POWER chips.
- For NVIDIA Pascal GPUs, we use the NVIDIA System Management Interface (nvidia-smi) for power measurements.

- Chips show 49%, memory units show 20%, GPUs show 18% variation in idle power consumption.
Dynamic Power Distribution

- We ran the micro-benchmarks independently on each processor to remove network variations.
- The power variation is 28% for DGEMM, 16% for KNeighbor, 20% for Stencil3D.
- Iso-performance processors: no significant performance variation (3%).
Idle and Active Power Correlation

- What metric should be used for sorting?
  - The chips that have high (or low) idle power do not necessarily have high (or low) active power.
  - Active power provides a better representation of the run-time scenario.
Temperature Distribution

• Would re-shuffling the hardware components cause temperature imbalance within data-center?

• Not significantly: Vertical distance is almost same as the horizontal distance.

• Cooling systems are designed for the worst case scenarios.
All Node Components Have Variation

• Distribution of the active power of different node components: CPU, GPU, Memory running DGEMM benchmark fit to the Gaussian distribution.
• Fitting curves are later used in evaluation for generating components for large-scale simulations.
Outline

1. Motivation
2. Power Variation Analysis
3. Variation Aware Node-Assembly Techniques
4. Evaluation
5. Conclusion
Variation Aware Node-Assembly Methods

Illustration of Type-1 Node Assembly

Illustration of Type-2 Node Assembly

Illustration of Type-3 Node Assembly

Sorted

Balanced

App-Aware
1. Sorted Assembly

- The goal is to sort the processors in terms of their power efficiency into nodes and racks.
- Place the most intensive workloads starting from the most efficient nodes.
- When the data center load is low, turn off inefficient nodes.
Data Center Utilization Varies Over Time

• Average weekly percentage utilization of different top supercomputers are shown during a period of seven months.
• Data is collected hourly starting Nov 1, 2017 from ANL, TACC and NERSC public websites.
• Avg utilization across all 5 supercomputers is 75%.
Power Reduction with Sorted Assembly

- Power reduction with sorted assembly compared to the random assembly at different data center loads with a size of 5,000 nodes.
- Unused nodes assumed to be turned-off.
2. Balanced Power Assembly

- The goal is to balance performance per watt for the nodes
2. Balanced Power Assembly

<table>
<thead>
<tr>
<th></th>
<th>CPU Variation</th>
<th>GPU Variation</th>
<th>Memory Variation</th>
<th>Min Node Power</th>
<th>Max Node Power</th>
<th>Node Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Assembly</td>
<td>27.8 %</td>
<td>18.3 %</td>
<td>21.5 %</td>
<td>1 (1097W)</td>
<td>1.15 (1267W)</td>
<td>14.4 %</td>
</tr>
<tr>
<td>Balanced Power Assembly</td>
<td>1.4 %</td>
<td>0.7 %</td>
<td>1.4 %</td>
<td>1.06 (1173W)</td>
<td>1.07 (1178W)</td>
<td>0.4 %</td>
</tr>
</tbody>
</table>

- Node to node power variation is minimized with power balanced assembly.
- Performance-per-watt becomes more predictable for nodes.
- This technique might be more suitable for cloud workloads.
3. Application Aware Assembly

• Components which application use most heavily are selected to use the most power efficient components.
• Job scheduler support is needed to decide application placement.
3. Application Aware Assembly

- With application-aware assembly:
  - CPU-intensive benchmarks run on the most power efficient half the CPUs, and inefficient part of GPUs.
  - GPU-intensive applications run on the most power efficient GPUs and inefficient half of the CPUs.
3. Application Aware Assembly

- Power reduction with application-aware assembly compared to random assembly.
- Power is normalized according to random assembly in each column.

- In a data-center comprised of 5,000 nodes, 2% of the node power is equivalent to 130 KW.
Outline

1. Motivation
2. Power Variation Analysis
3. Variation Aware Node-Assembly Techniques

4. Evaluation

5. Conclusion
Evaluation – $ Savings

$AC = Additional Assembly Cost
$ER = Energy Reduction
$PR = Power Reduction
$EP = Electricity Price = 10.48 cents per kWh [26]
$T = System Up Time = 350 days = 8400 hours
Average System Utilization = 50% (or 3.5% total reduction)

$CostReduction = EP \times ER - AC
0 < EP \times T \times PR - AC
AC < 10.48 (cents per kWh) \times 8400 hours \times 100 KW
AC < $90,083 per year

• Dollar savings increase as the data center size increases.
What if variation increase?

- Power reduction increases as variability increases for sorted assembly.
- \( \sigma \) represents measured standard deviation in the current architectures. \( 1.5\sigma, 2\sigma \) represents the scenarios when the deviation increases 1.5x and 2x times respectively.
Outline

1. Motivation
2. Power Variation Analysis
3. Variation Aware Node-Assembly Techniques
4. Evaluation
5. Conclusion
Summary

- There is significant manufacturing variation among components in HPC data-centers
- Node assembly techniques do not take hardware variation into account
- We propose and evaluate three node assembly techniques

<table>
<thead>
<tr>
<th></th>
<th>Use Cases</th>
<th>Job Scheduler Support</th>
<th>Energy Savings</th>
<th>Performance Degradation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorted Assembly</td>
<td>Systems with variable utilization rates</td>
<td>Minor</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Balanced Power Assembly</td>
<td>Cloud systems</td>
<td>None</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>App. Aware Assembly</td>
<td>Systems with mixed app. characteristics</td>
<td>Major</td>
<td>✓</td>
<td>×</td>
</tr>
</tbody>
</table>
Thank you!
Backup Slides: Power 8 & 9 Parametric Data
Backup Slides: Performance Variation

**Idle Chip Power and IPS**

- **Graph 1:** Scatter plot showing the relationship between idle chip power (normalized to $\mu_{c\text{-idle}}$) and IPS (Millions).
- **Graph 2:** Scatter plot showing the relationship between chip power (normalized to $\mu_{c\text{-idle}}$) and IPS (Millions).

*Note: Data points represent P8 Chip performance.*