Power Aware Heterogeneous Node Assembly

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

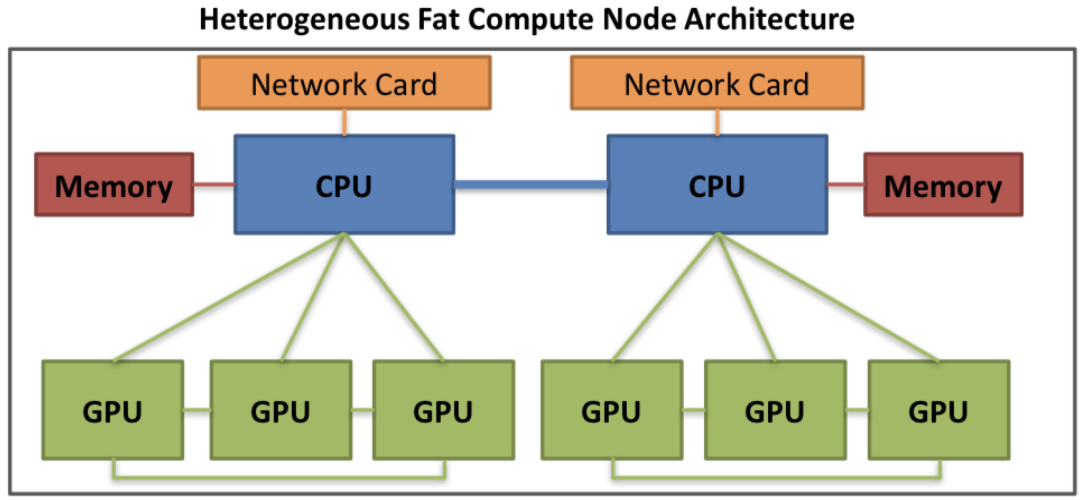
Outline

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

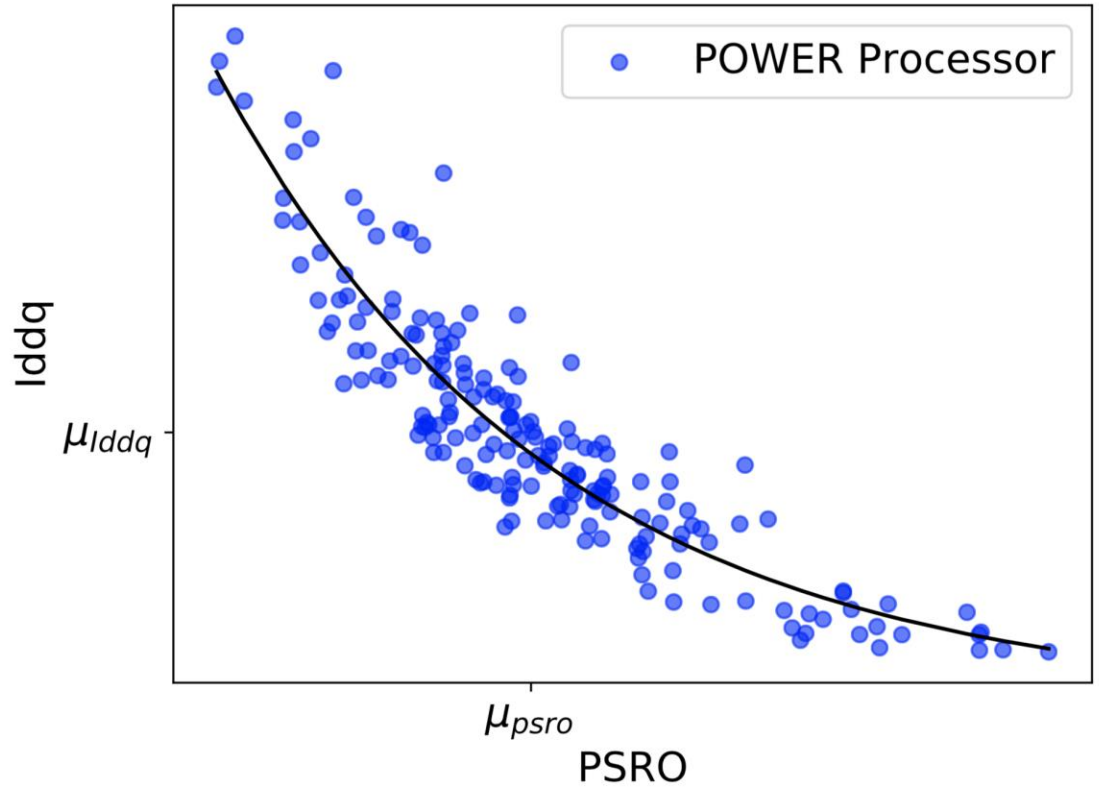
Motivation: Heterogenous Fat Compute Nodes

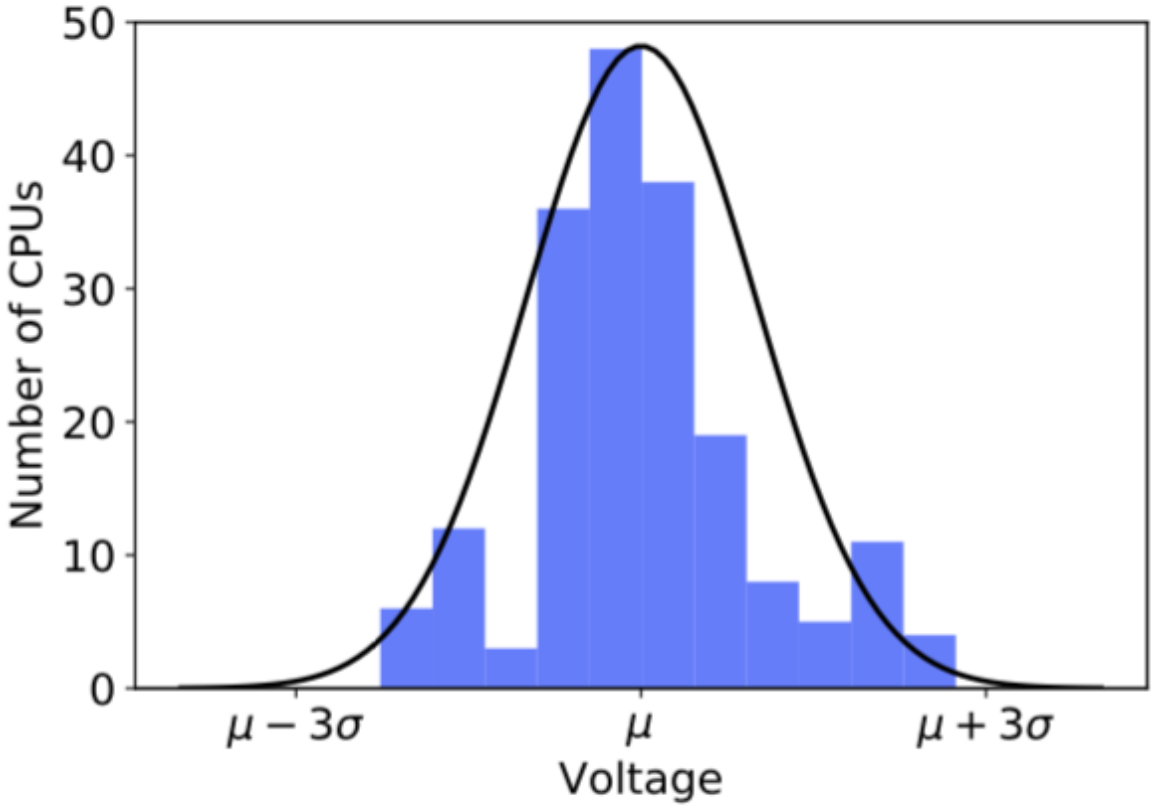


www.top500.org
November, 2018 list

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

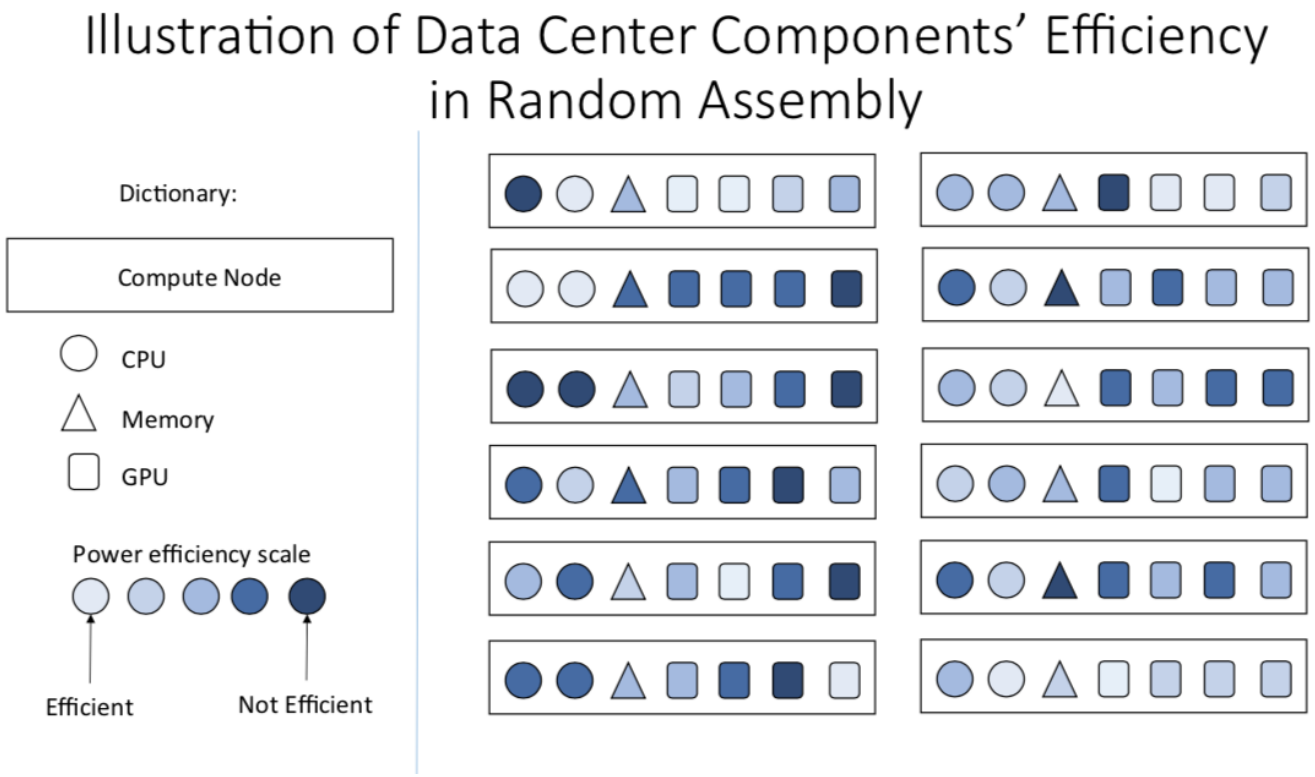
Motivation: Manufacturing Variations in Hardware





- 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.

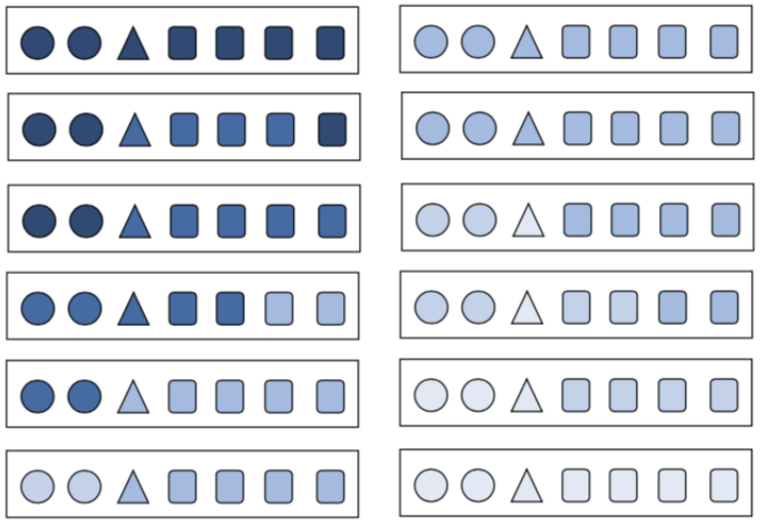
Motivation: Insufficient Scheduling Methods



- 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

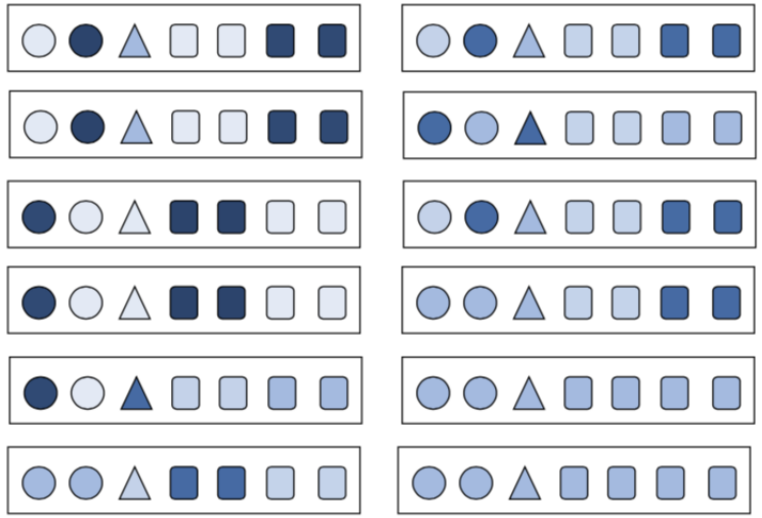
Variation Aware Node-Assembly Methods

Illustration of Type-1 Node Assembly



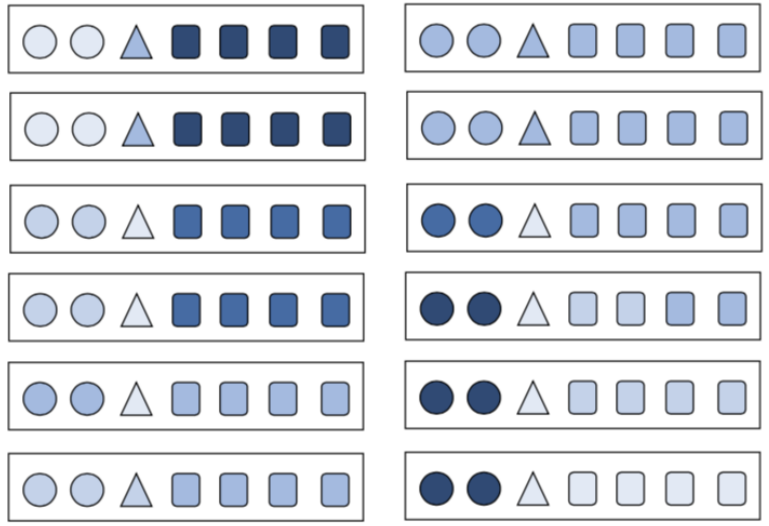
Sorted

Illustration of Type-2 Node Assembly



Balanced

Illustration of Type-3 Node Assembly



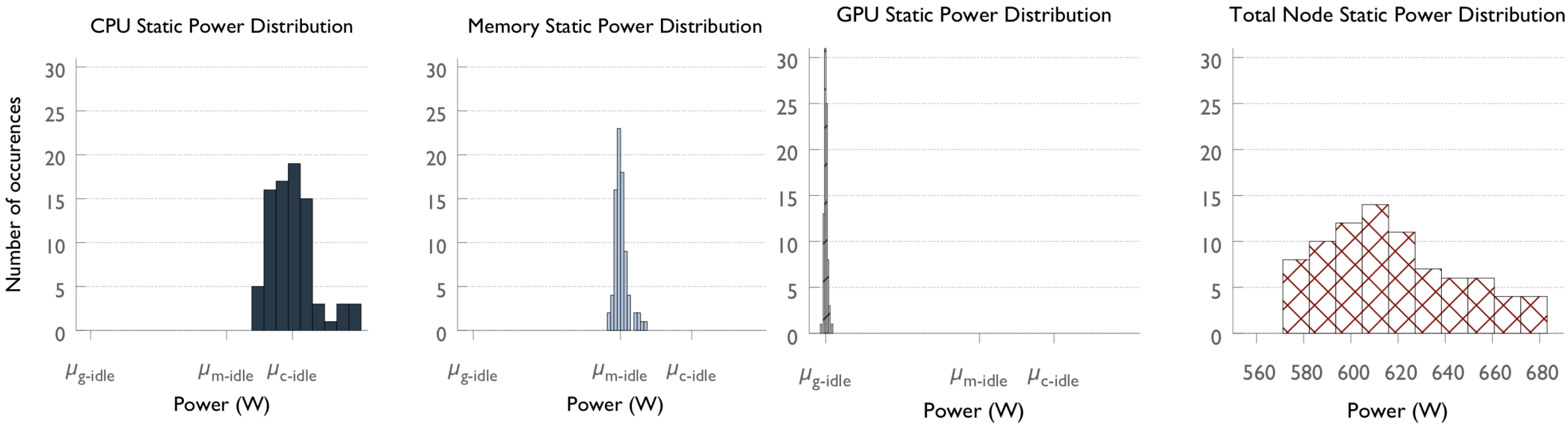
App-Aware

Outline

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

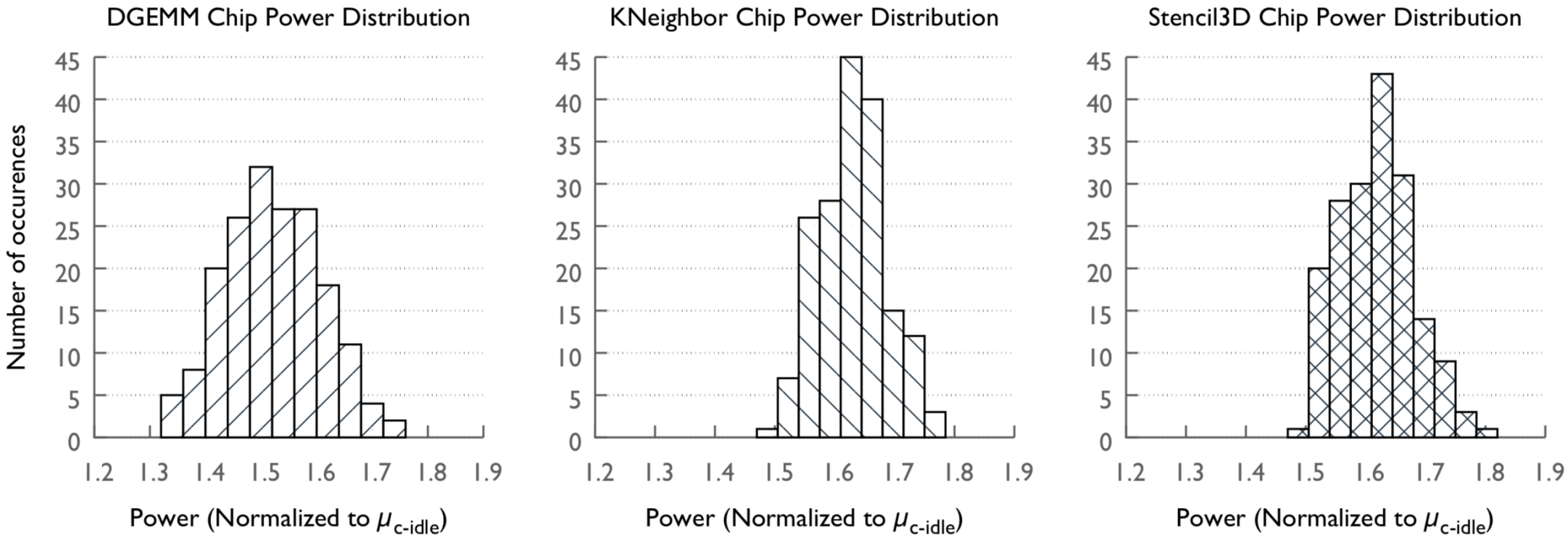
Static Power Distribution

- 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.



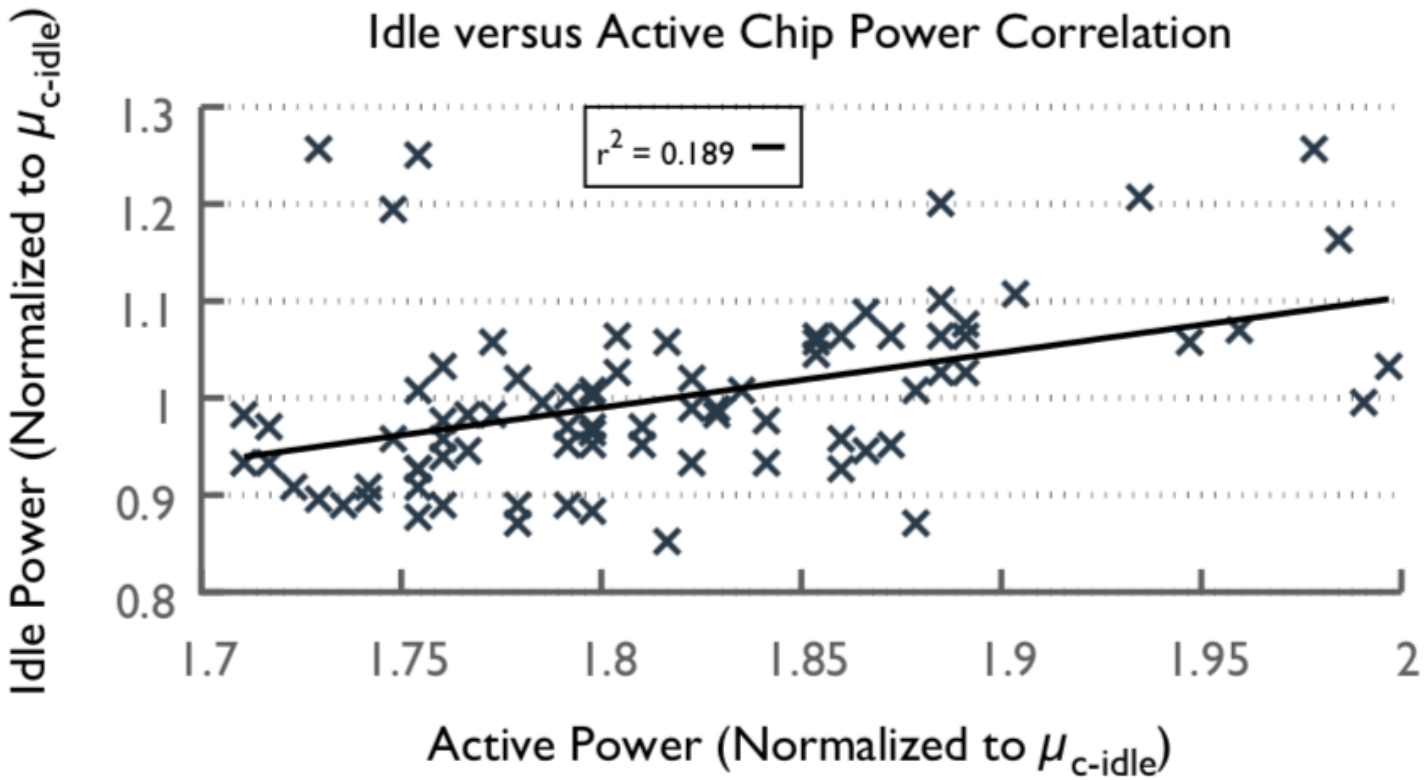
- Chips show 49%, memory units show 20%, GPUs show 18% variation in idle power consumption.

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%).

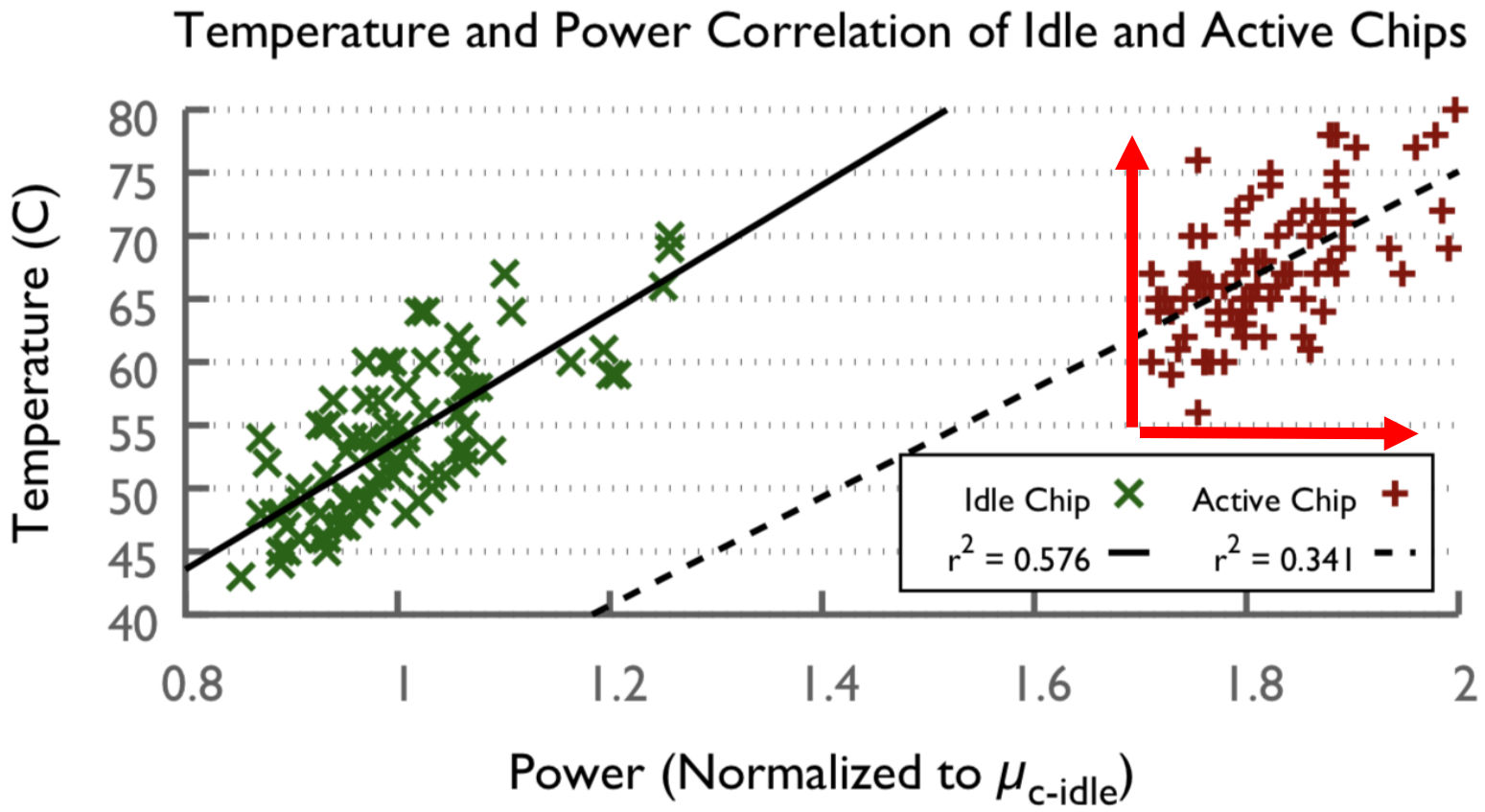
Idle and Active Power Correlation



- 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.

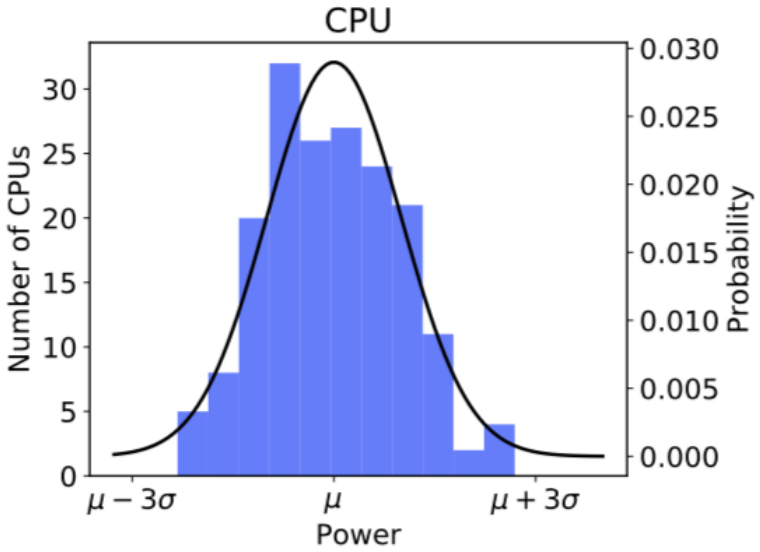
Temperature Distribution

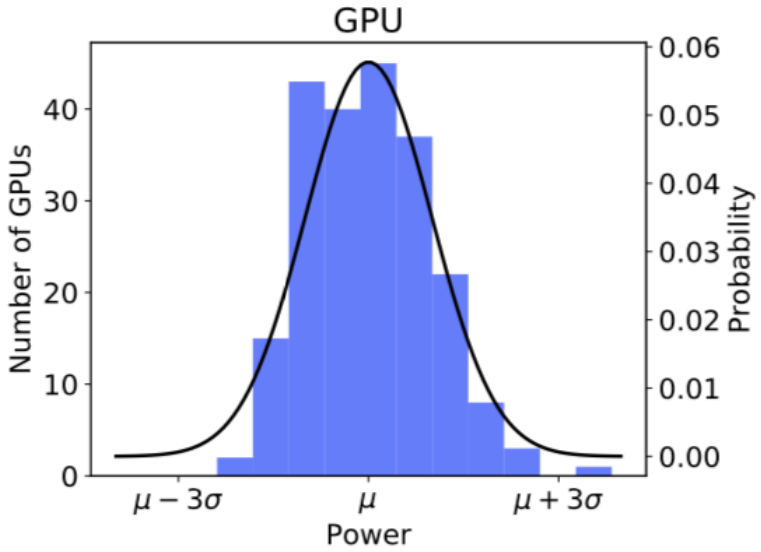
- Would re-shuffling the hardware components cause temperature imbalance within data-center?

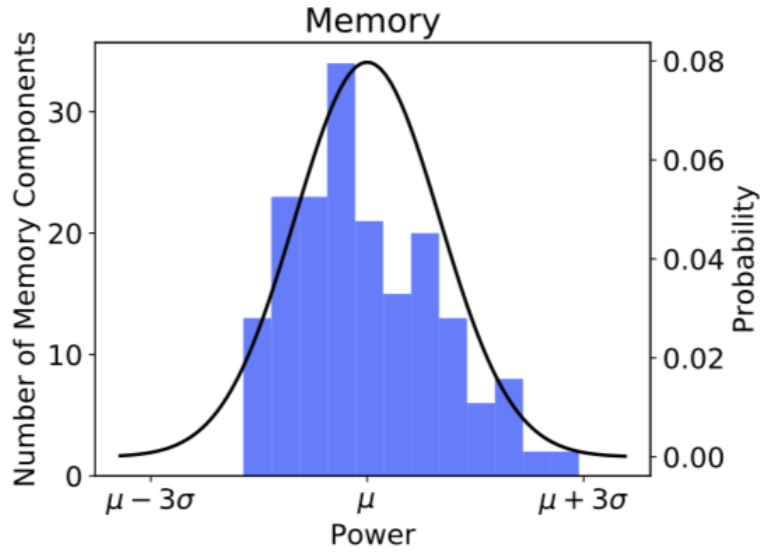


- **Not significantly:** Vertical distance is almost same as the horizontal distance.
- Cooling systems are designed for the worst case scenarios.

All Node Components Have Variation







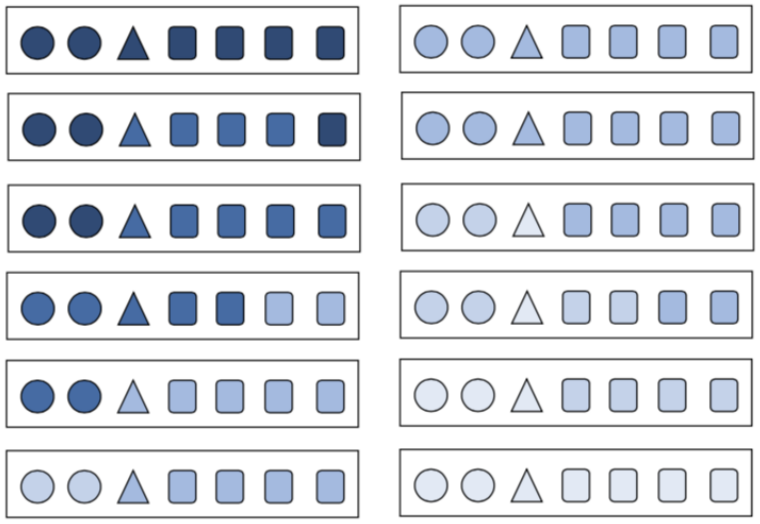
- Distribution of the active power of different node components: CPU, GPU, Memory running DGEMM benchmark fit to the Gaussian distribution.
- Fitting curves are later used in evaluation for generating components for large-scale simulations.

Outline

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

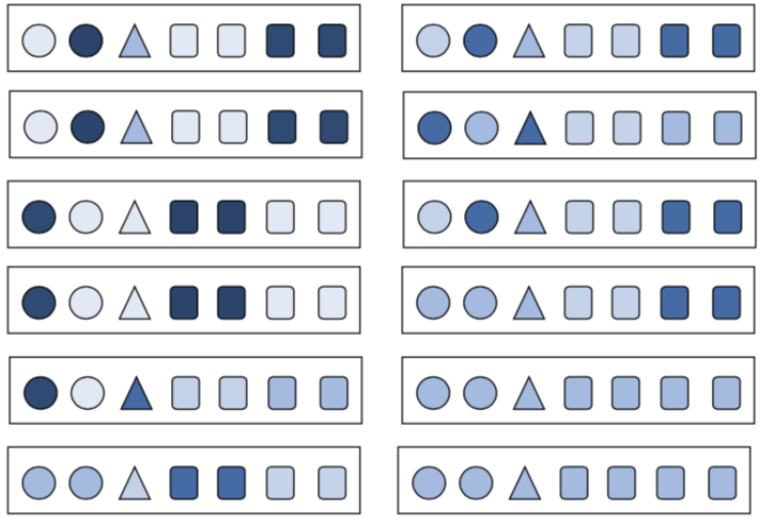
Variation Aware Node-Assembly Methods

Illustration of Type-1 Node Assembly



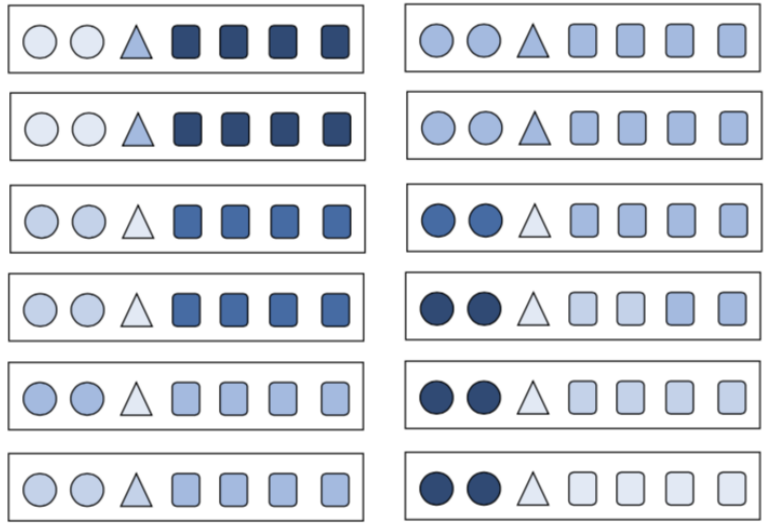
Sorted

Illustration of Type-2 Node Assembly



Balanced

Illustration of Type-3 Node Assembly



App-Aware

1. Sorted Assembly

- The goal is the 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 in-efficient nodes

Illustration of Data Center Components' Efficiency in Random Assembly

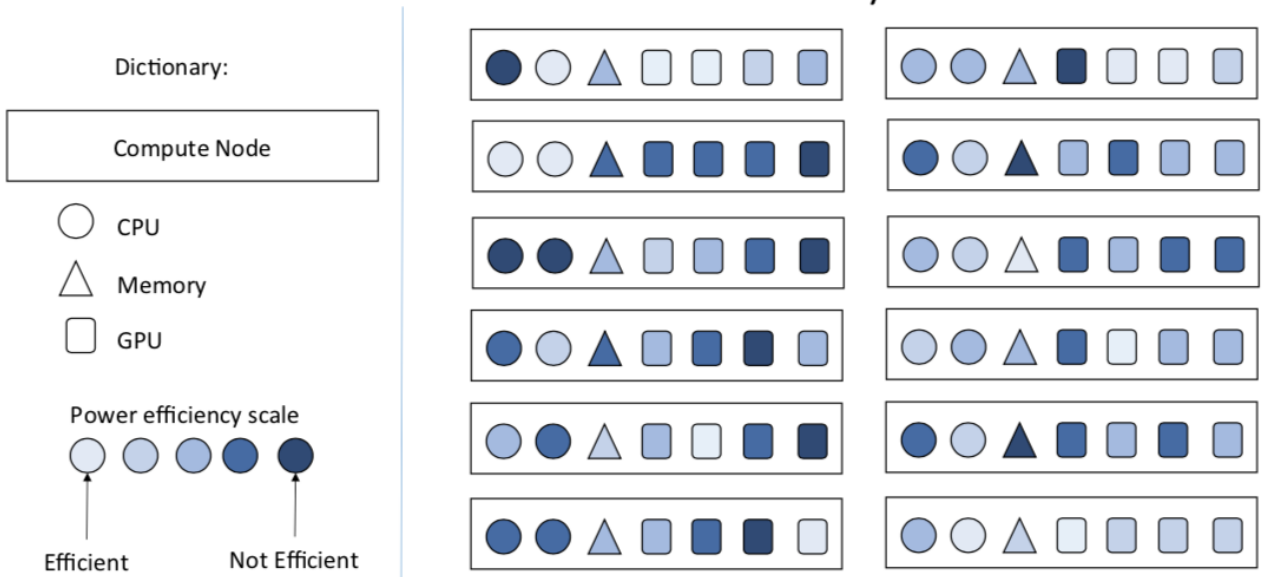
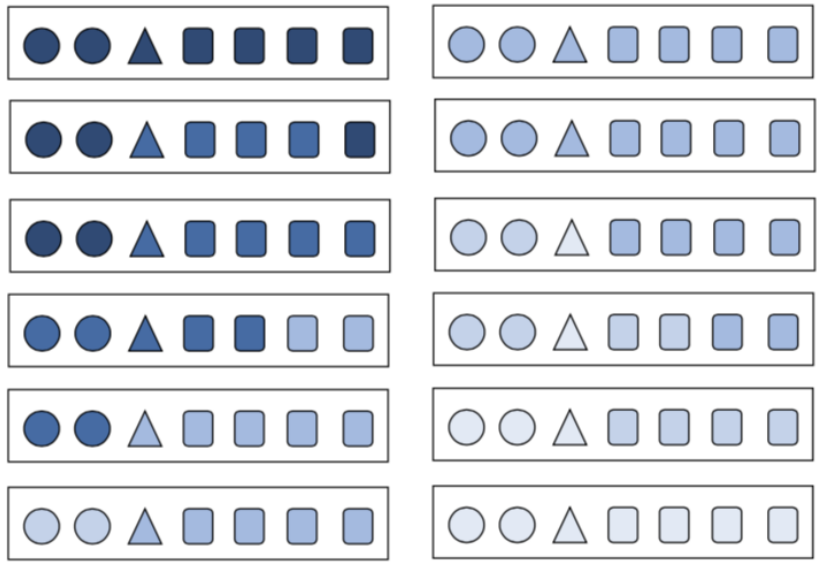
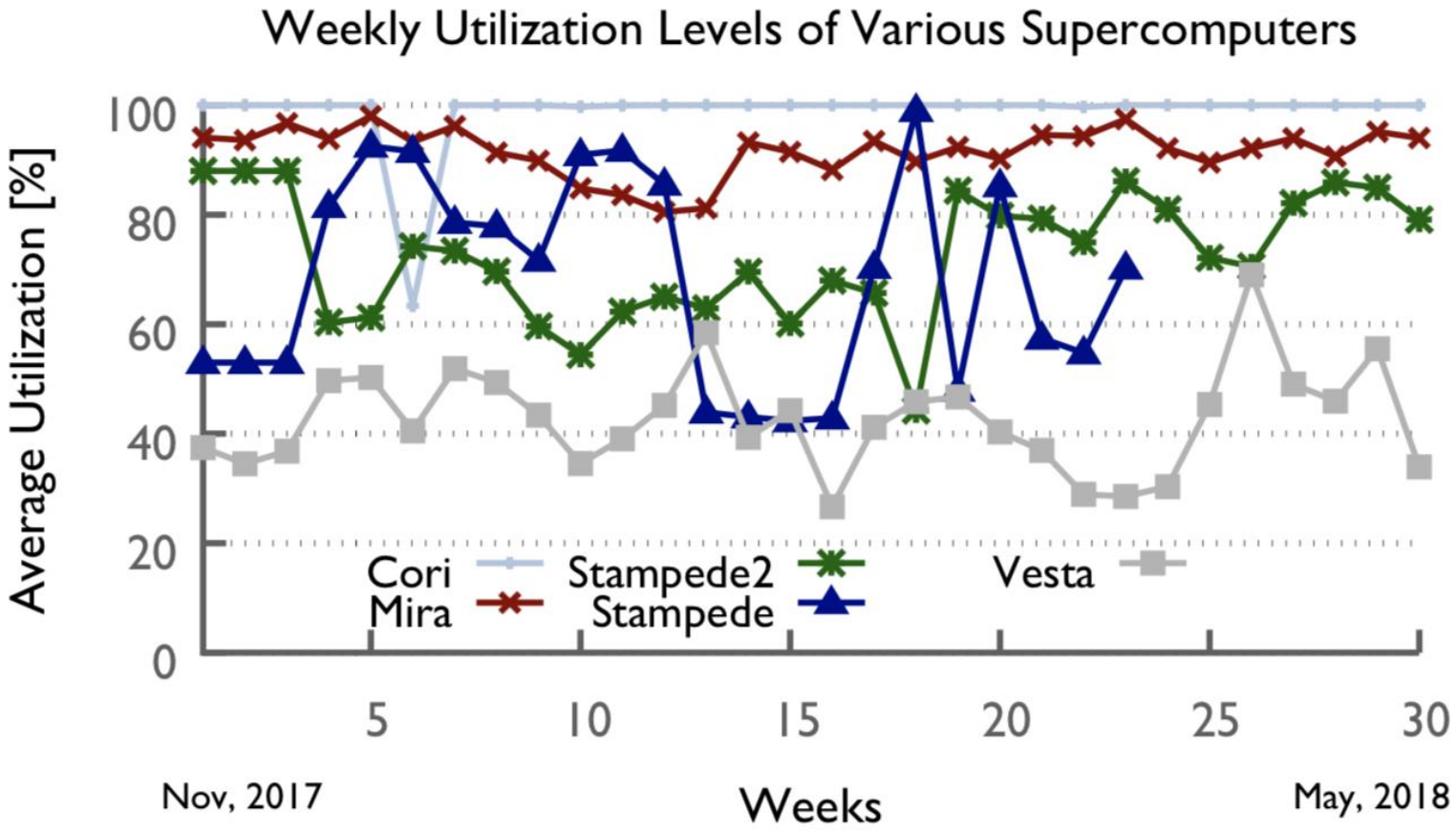


Illustration of Type-1 Node Assembly

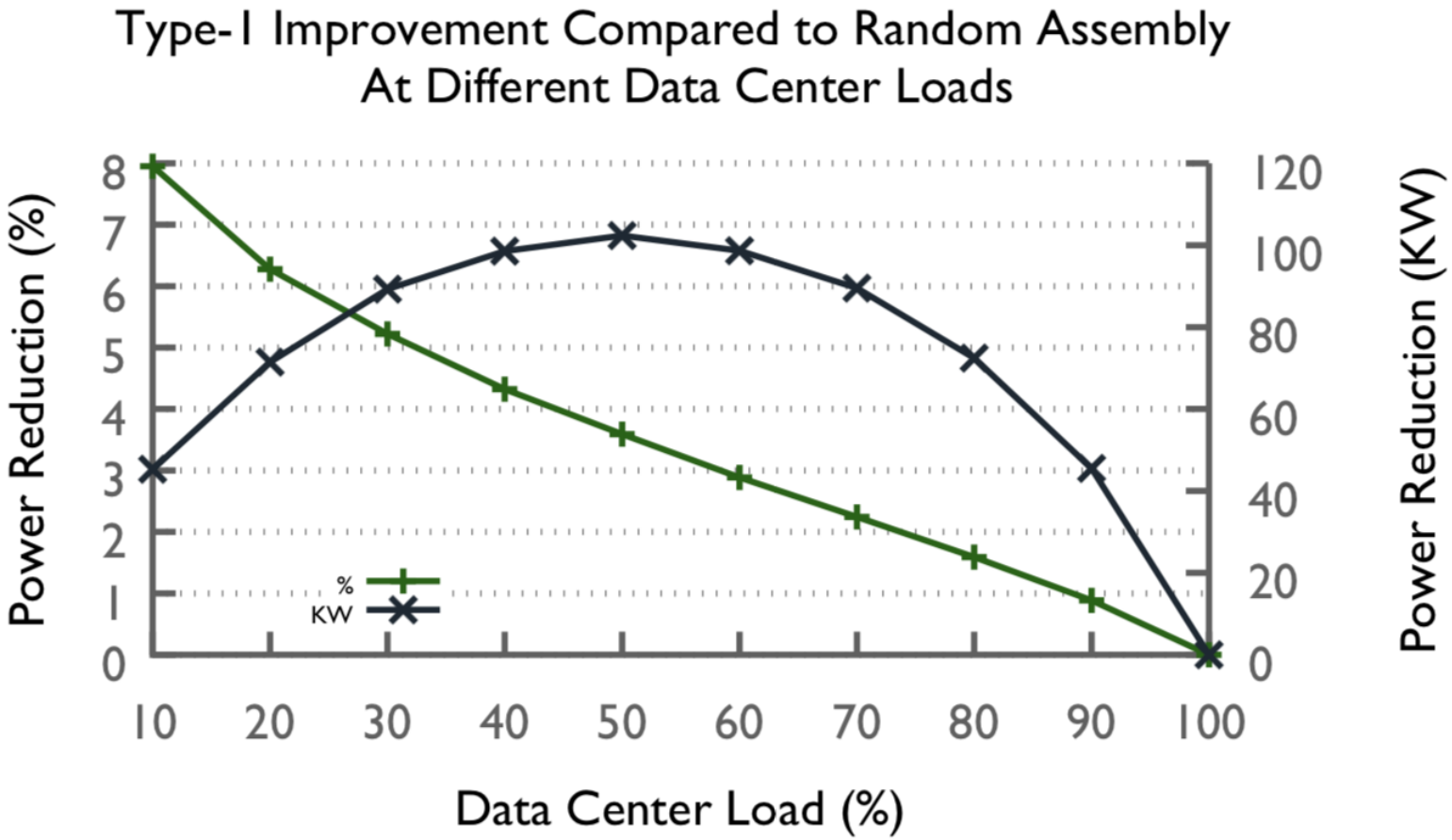


Data Center Utilization Varies Over Time



- 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

Illustration of Data Center Components' Efficiency in Random Assembly

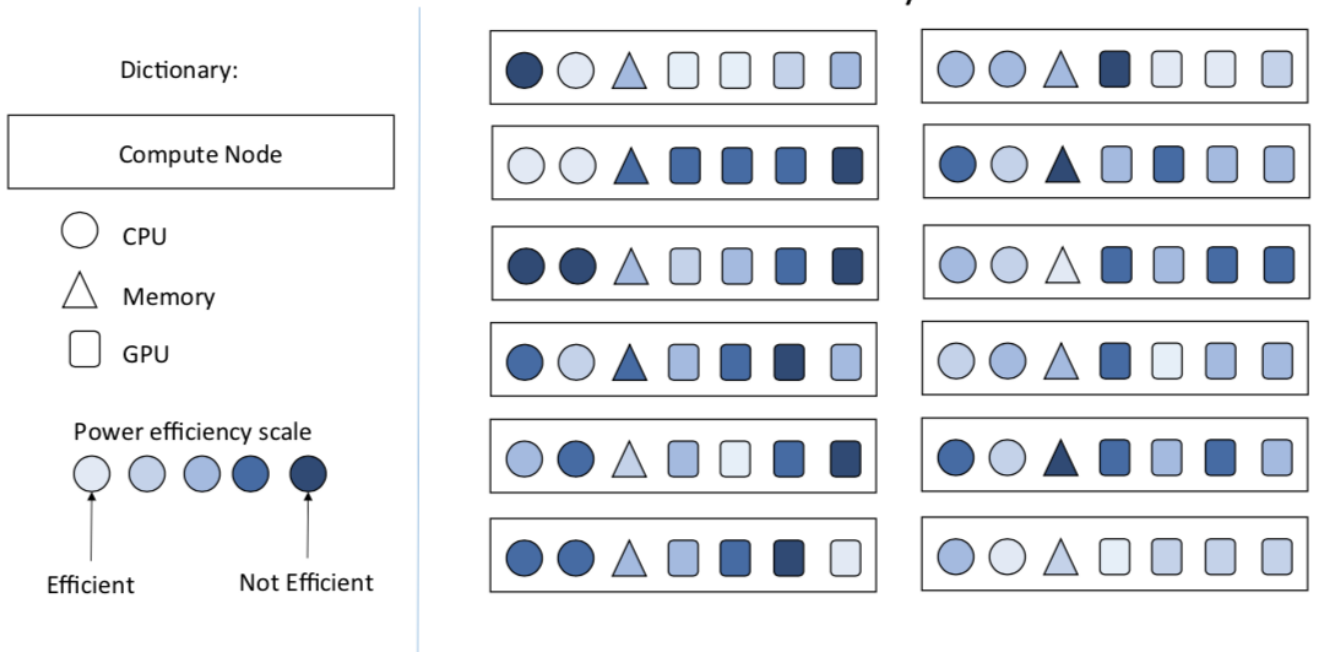
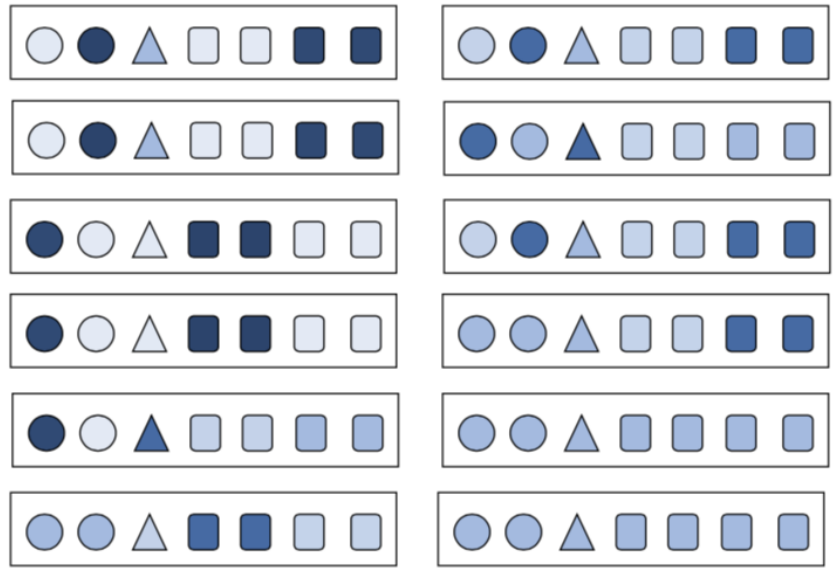


Illustration of Type-2 Node Assembly



2. Balanced Power Assembly

	CPU Variation	GPU Variation	Memory Variation	Min Node Power	Max Node Power	Node Variation
Random Assembly	27.8 %	18.3 %	21.5 %	1 (1097W)	1.15 (1267W)	14.4 %
Balanced Power Assembly	1.4 %	0.7 %	1.4 %	1.06 (1173W)	1.07 (1178W)	0.4 %

- 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.

Illustration of Data Center Components' Efficiency in Random Assembly

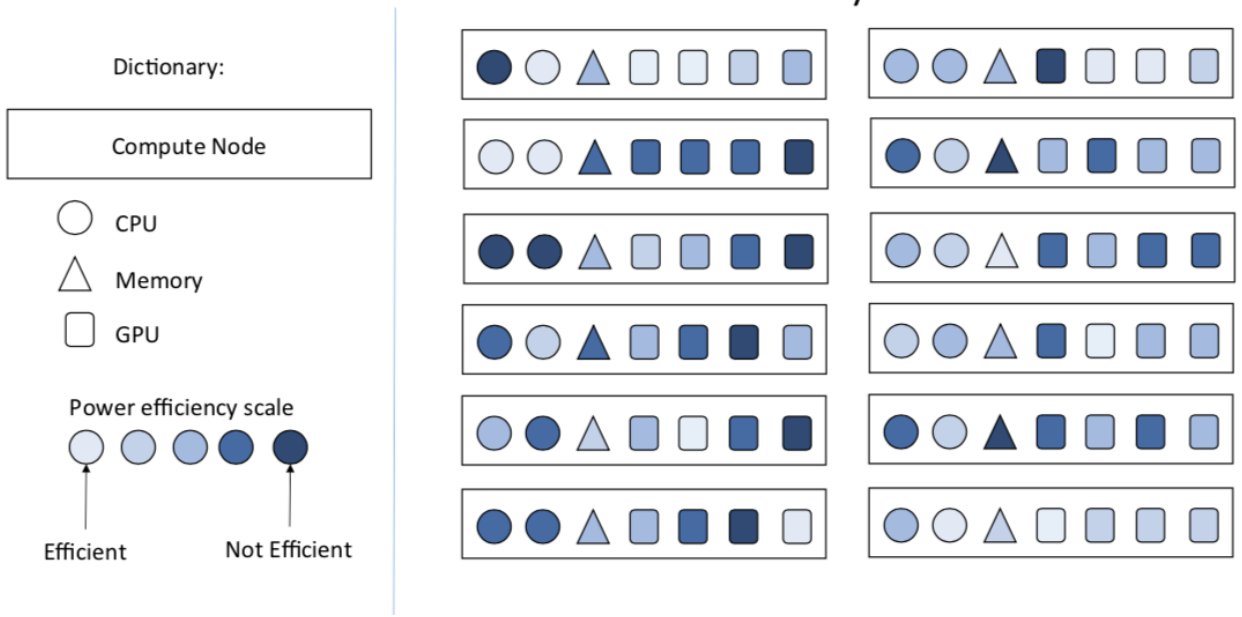
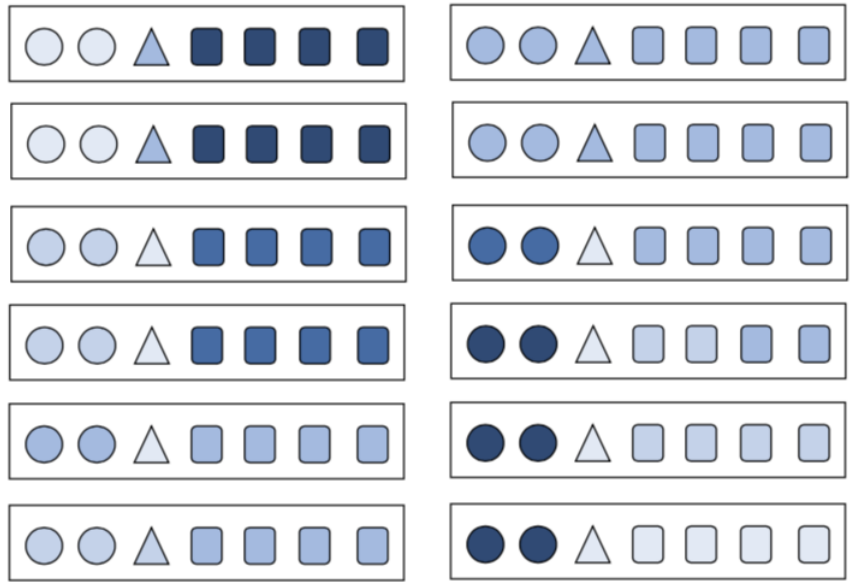
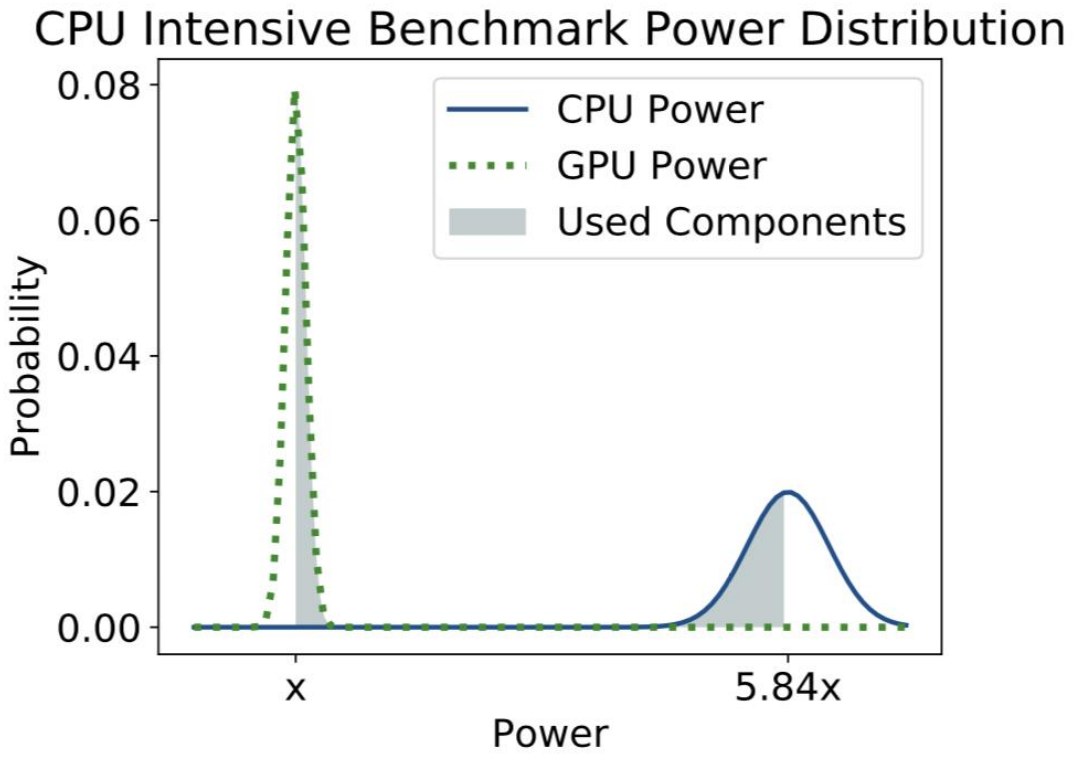
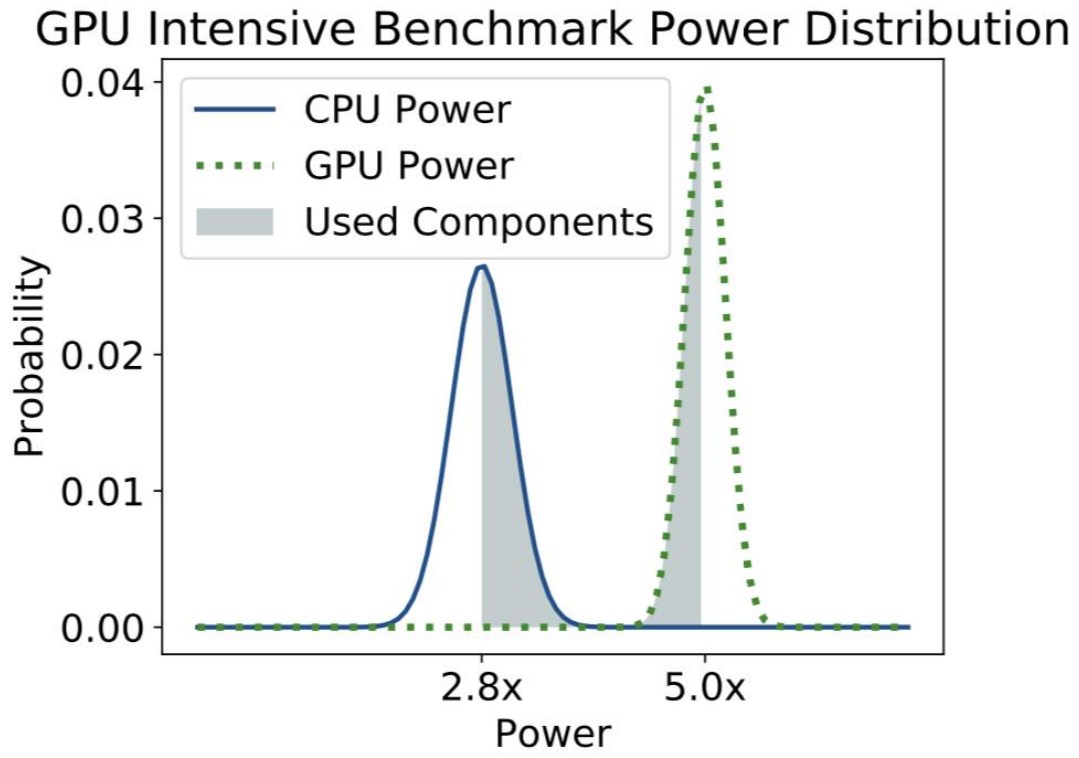


Illustration of Type-3 Node Assembly



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

	CPU Power	GPU Power	Node Power
Random Assembly	1	1	1
App. Aware Assembly	0.97	0.98	0.98
Power Reduction	3%	2%	2%

- 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

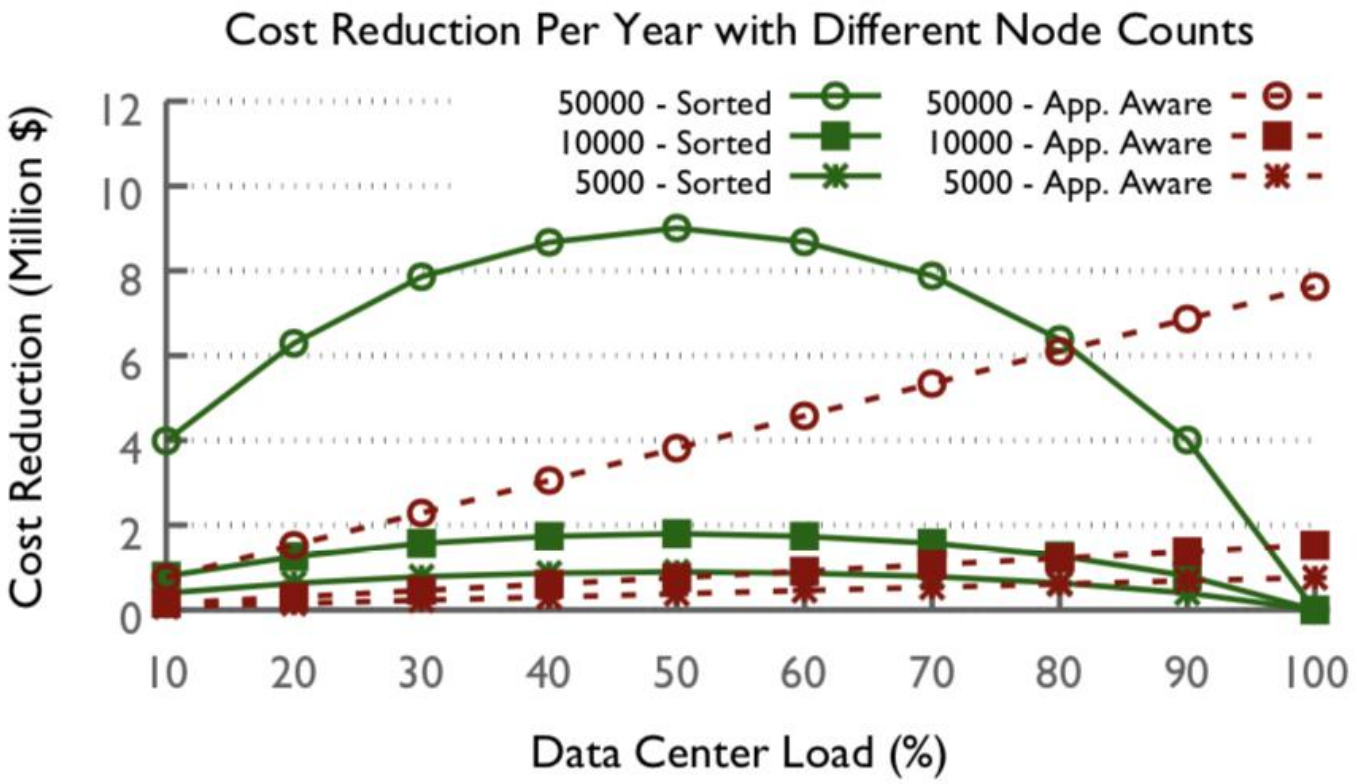
AC = *Additional Assembly Cost*
ER = *Energy Reduction*
PR = *Power Reduction*
EP = *Electricity Price* = *10.48 cents per kWh* [26]
T = *System Up Time* = *350 days* = *8400 hours*
Average System Utilization = *50%* (*or 3.5% total reduction*)

$$CostReduction = EP \times ER - AC$$

$$0 < EP \times T \times PR - AC$$

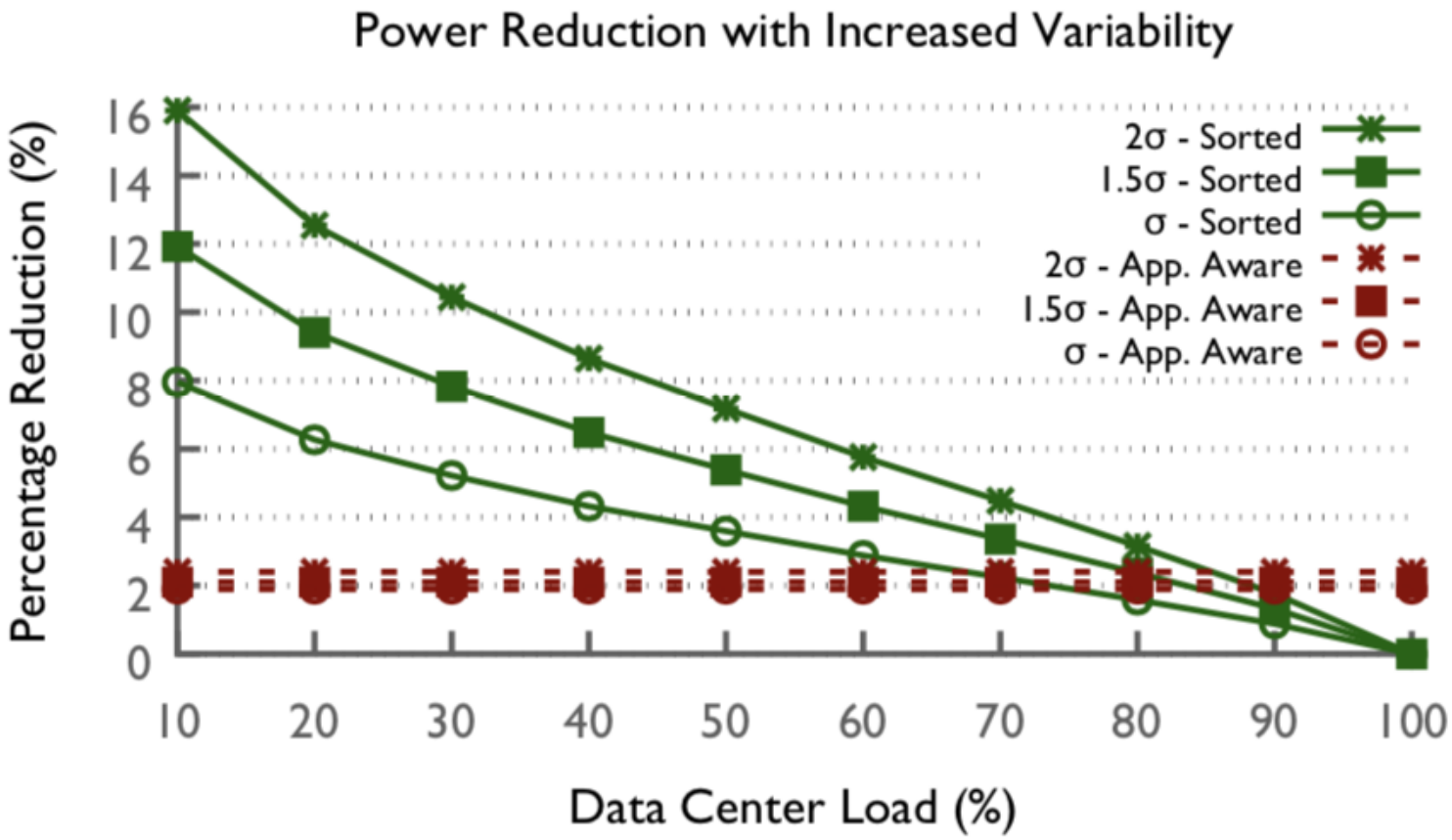
$$AC < 10.48\ (cents\ per\ kWh) \times 8400\ hours \times 100\ KW$$

$$AC < \$90,083\ per\ year$$



- Dollar savings increase as the data center size increases.

What if variation increase?



- Power reduction increases as variability increases for sorted assembly.
- σ represents measured standard deviation in the current architectures. 1.5σ, 2σ 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**

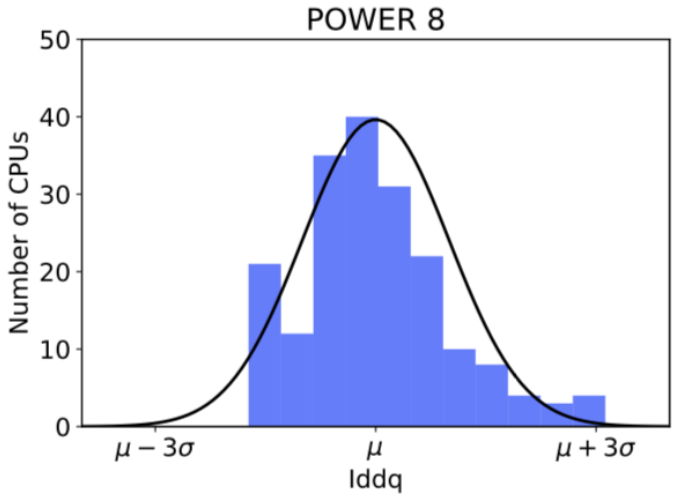
Summary

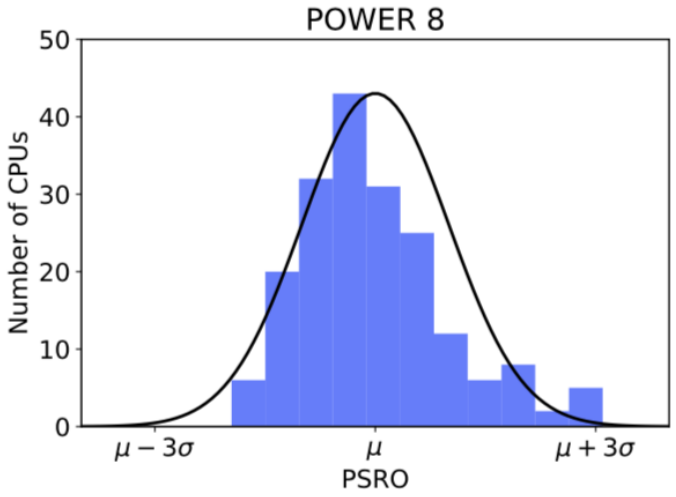
- There is significant manufacturing variation among components in HPC data-centers
- Node assembly techniques do not take hardware variation into account
- We propose and evaluate three node assembly techniques

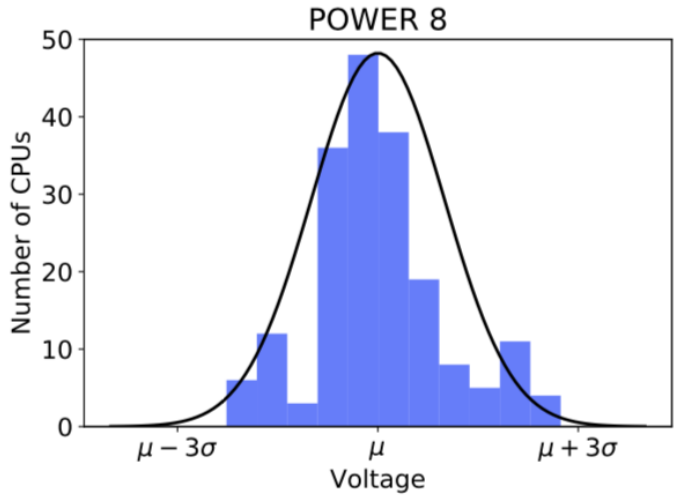
	Use Cases	Job Scheduler Support	Energy Savings	Performance Degradation
Sorted Assembly	Systems with variable utilization rates	Minor	✓	×
Balanced Power Assembly	Cloud systems	None	×	×
App. Aware Assembly	Systems with mixed app. characteristics	Major	✓	×

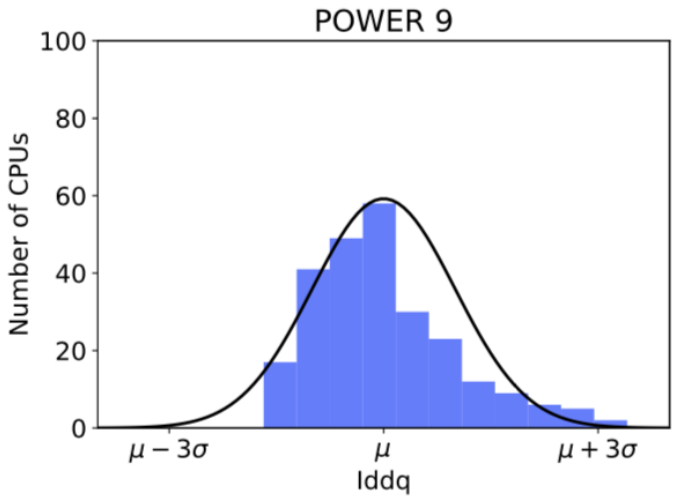
Thank you!

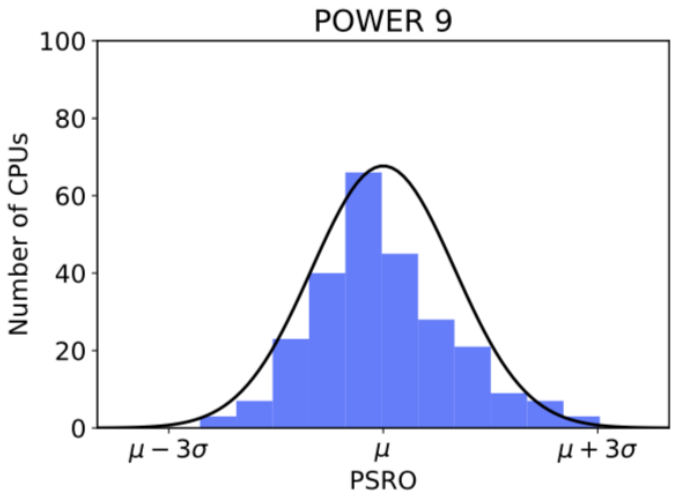
Backup Slides: Power 8 & 9 Parametric Data

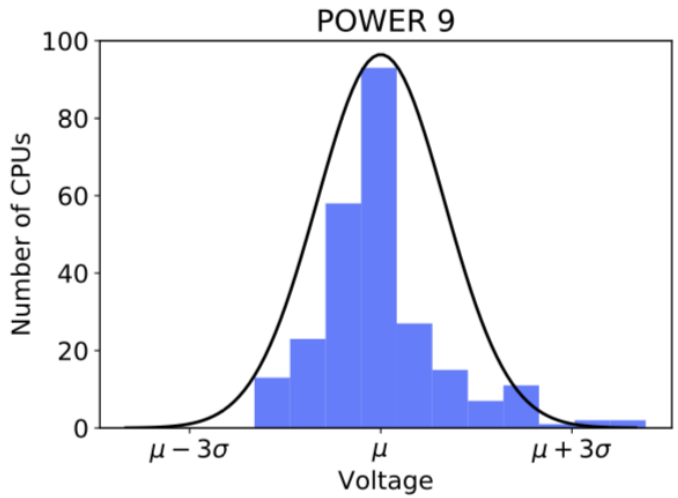












Backup Slides: Performance Variation