A Holistic View of Memory Utilization on Perlmutter

Jie Li[1], George Michelogiannakis[2], Brandon Cook[2] , Yong Chen[1]
[1]Texas Tech University, [2]Lawrence Berkeley National Laboratory

ABSTRACT

HPC systems are at risk of being underutilized due to various resource requirements of applications and the imbalance of utilization among subsystems. This work provides a holistic analysis and view of memory utilization on a leadership computing facility, the Perlmutter system at NERSC, through which we gain insights about the resource usage patterns of the memory subsystem. The results of the analysis can help evaluate current system configurations, offer recommendations for future procurement, provide feedback to users on code efficiency, and motivate research in new architecture and system designs.

BACKGROUND

Perlmutter[1]:

- Ranked 7th in the top500 list[2].
- More than 1,500 GPU nodes and over 3,000 CPU nodes.
- GPU node : four NVIDIA A100 Tensor Core GPUs, one AMD "Milan" CPU, 160GB of HBM2, and 256GB of DRAM.
- CPU node: two AMD "Milan" CPUs and 512GB of DRAM.

Data Collection:

- LDMS collects system-level metrics on GPU and CPU nodes; DCGM collects GPU metrics.
- Metrics are collected from June 15 to July 1, 2022 at an interval of 10 seconds and are saved in CSV files.
- *LDMS_ETL* joins CSV files with SLURM *sacct* data and saves merged metrics including Job ID and Job Steps info in parquet files.

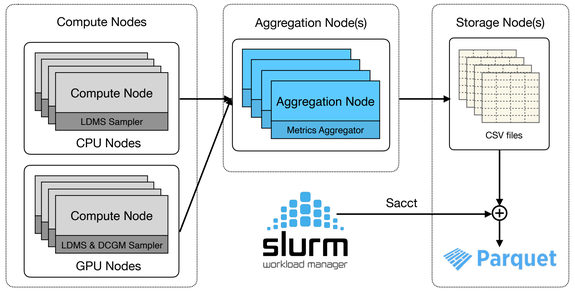

Figure 1. Workflow of Monitoring Data Collection

JOB SIZE DISTRIBUTION

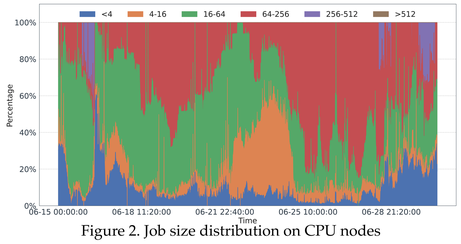

Figure 2. Job size distribution on CPU nodes

Table 1: Average Percentage of Each Job Group

	< 4	4-16	16-64	64-256	256-512	> 512
CPU nodes	9.89%	11.34%	37.89%	37.92%	2.96%	0%
GPU nodes	27.29%	10.09%	11.46%	40.91%	3.26%	6.99%

JOB SIZE DISTRIBUTION (Cont.)

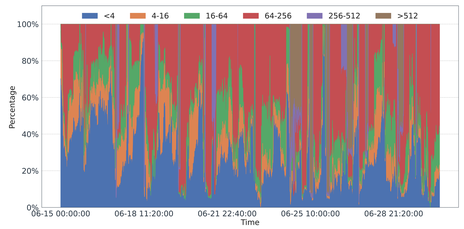

Figure 3. Job size distribution on GPU nodes

Observation:

- CPU nodes have more middle-scale (4-64 nodes) jobs.
- GPU nodes have more small-scale jobs (<4 nodes) and large-scale jobs (>64 nodes).

Conclusion:

- Small-scale jobs on GPU nodes could be attributed the emerging ML/DL applications.
- The extremely large-scale jobs prefer to exploit GPU nodes to achieve a faster simulation or training process.

NODE-LEVEL MEMORY UTILIZATION

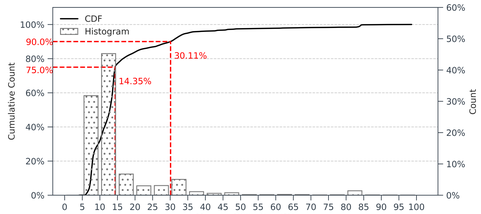

Figure 4. DRAM utilization on CPU nodes

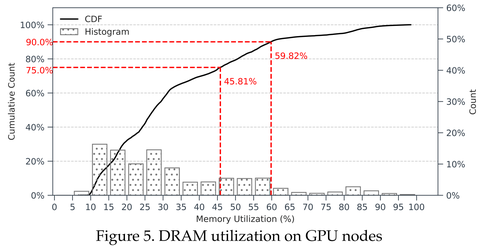

Figure 5. DRAM utilization on GPU nodes

Observation:

- CPU nodes have their 90th percentile in DRAM utilization at 30.11%, indicating that less than 154 GB memory is used for 90% of the time.
- Almost 60% of DRAM on GPU nodes is used for 90% of the time, corresponding to 154 GB.

Conclusion:

- The long tail of the CPU CDF indicates that the chances of a CPU node to exhaust its memory resource are low.
- The slowly increase of CDF on GPU nodes indicates that GPU nodes have a more balanced DRAM utilization than that on CPU nodes.

JOB-LEVEL MEMORY UTILIZATION

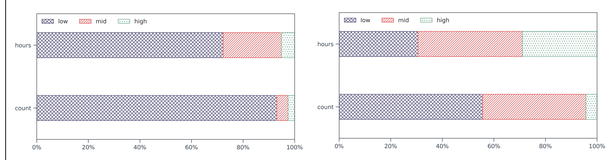

Figure 6. Decomposition of the memory intensity in hours and number of CPU jobs (left) and GPU jobs (right).
(DRAM Memory Intensity: Low: <25%, Mid: 25-50%, High: >50%)

Observation:

- About 93% of CPU jobs use less than 25% of the total memory capacity.
- Moderate and high memory intensity CPU jobs only take up 7% of the total jobs but consume about 28% of the node-hours; 44% of DRAM intensive GPU jobs account for 68% of the node-hours.

Conclusion:

- Moderate and high memory intensity jobs likely use more nodes and/or run longer time.
- Most of CPU jobs can be accommodated with reduced memory capacity nodes.

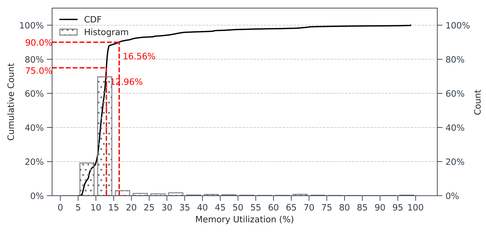

Figure 7. Maximum DRAM utilization of CPU jobs

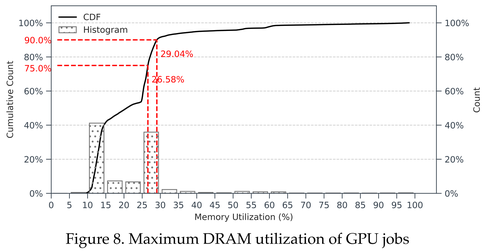

Figure 8. Maximum DRAM utilization of GPU jobs

Observation:

- 90% of CPU jobs only use up to 16.59% of total DRAM capacity, corresponding to 85 GB.
- For 90% of the GPU jobs, their maximum DRAM usage is no larger than 29.04% of total memory capacity, i.e., 74 GB.

Conclusion:

- The long tail of the CDF plot indicates that very few jobs will take up all the DRAM capacity on CPU and GPU nodes.
- DRAM on both CPU and GPU nodes can be reduced to under 85 GB without interfering 90% of jobs.

GPU MEMORY (HBM2) UTILIZATION

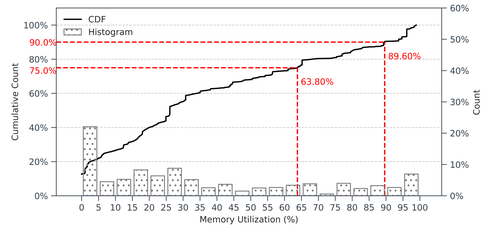

Figure 9. HBM2 utilization on GPU nodes

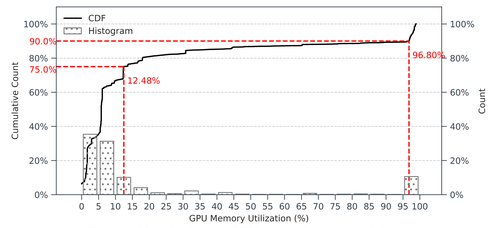

Figure 10. Maximum HBM2 utilization of jobs

Observation:

- For 90% of the time, the HBM2 utilization is nearly 90%.
- 75% of jobs have the maximum HBM2 utilization less than 12.48%, i.e. about 20 GB.
- About 10% of jobs use over 95% of the total HBM2 capacity.

Conclusion:

- GPU HBM2 have a relatively balanced utilization across all time.
- Only a small fraction of jobs can take full advantage of HBM2.

SUMMARY AND FUTURE WORK

Summary:

- Memory resources are under-utilized/over-provisioned both on CPU nodes and GPU nodes.
- Most of GPU jobs cannot use HBM2 resources effectively.

Future work:

- Analyzing the memory resources in temporal and in spatial.
- Extending the analysis on other subsystems.

REFERENCES

[1] NERSC. (2022) Perlmutter. [Online]. Available: https://www.nersc.gov/systems/perlmutter/
[2] Top500. (2022) Top500 list. [Online]. Available: https://www.top500.org/lists/top500/2022/06/
[3] A. Agelastos, B. Allan, J. Brandt, P. Cassella, J. Enos, J. Fullop,A. Gentile, S. Monk, N. Naksinehaboon, J. Ogden et al., "The lightweight distributed metric service: a scalable infrastructure for continuous monitoring of large scale computing systems and applications," in SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2014, pp. 154–165.
[4] NVIDA. (2022) NVIDIA DCGM. [Online]. Available: https://developer.nvidia.com/dcgm
[5] J. Peng, J. Karlin, M. Gokhale, K. Shoga, M. Legendre, and T. Gamblin, "A holistic view of memory utilization on hpc systems: Current and future trends," in The International Symposium on Memory Systems, 2021, pp. 1–11.
[6] G. Michelogiannakis, B. Klenk, B. Cook, M. Y. Teh, M. Glick, L. Dennison, K. Bergman, and J. Shalf, "A case for intra-rack resource disaggregation in hpc," ACM Transactions on Architecture and Code Optimization (TACO), vol. 19, no. 2, pp. 1–26, 2022.

ACKNOWLEDGMENTS

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.