

The Neuronix HPC Cluster: Cluster Management Using Free and Open Source Software Tools¹

C. Campbell, N. Mecca, I. Obeid and J. Picone

The Neural Engineering Data Consortium, Temple University
{christopher.campbell, tuf89560, iobeid, picone}@temple.edu

One of the most notable impacts of computing advancements over the last few decades has been the decentralization of resources. As the cost of computer hardware continues to decrease, significant computational power continues to become more accessible to the consumer market. Similar to personal computing, this phenomenon has also enabled the growth of low-cost high-performance computing (HPC) (e.g., desktop supercomputers). Combined with advances in computational statistics and machine learning, HPC systems can now accommodate computationally expensive research using consumer-grade hardware.

Graphic Processing Units (GPUs) have become an integral part of today's high-performance compute cluster [1][2]. GPUs are absolutely critical now to a new generation of big data machine learning systems that require massive amounts of computing to develop. These chips and the software that supports them adds additional burden to cluster management. Key issues include software compatibility (e.g., Nvidia's CUDA support is problematic), as well as job control (e.g., open source schedulers do not seem to adequately support nodes with multiple GPU chips) and load balancing by distributing computing jobs to compute nodes based on the state they report to the main node. When there are large numbers of compute nodes in the cluster, system administration of these nodes becomes a time-consuming process. The goal of this poster presentation is to introduce researchers to cost-effective ways to manage such resources.

In the Neuronix cluster, we manage compute nodes by placing them under the control of the main server (i.e. CPU/GPU compute nodes). We use Warewulf [3] for operating system provisioning as well as for synchronizing important system files such as the hosts and password files. Warewulf boots the compute nodes over the network from kernel and filesystem images on the main server. The primary advantage of this architecture is that changes can be made to one set of images and sent to all the nodes. For the nodes that function independently of the main node (e.g. backup servers, web servers), we are in the process of implementing Ansible [4] to automate their setup and configuration.

The queue manager that controls job submission is composed of a resource manager (TORQUE [5]), monitors node resource statistics. It also handles everything related to submitting and running jobs from the main node on one or more compute nodes. Torque is accompanied by a job scheduler (Maui [6]) that communicates with the resource manager and, based on the status of the compute nodes and the internal scheduling of node can be requested, and setting scheduling single or multi-dimensional scheduling policies (e.g. setting scheduling policies per user per queue).

A number of free and open source monitoring tools are available to keep track of system statistics (e.g. network bandwidth, CPU/memory usage) and hardware failures. As computing systems scale, it is important to identify and resolve bottlenecks, which can limit the performance gain from scaling. Ganglia [7] is a system monitoring software that collects information from cluster nodes and displays the information graphically through a web interface [8]. *Mdadm* is a standard Linux utility that we use to manage the RAID arrays on the cluster. With these arrays, the cluster becomes significantly more robust in the face of hard drive failures [9]. *Smartctl* is another standard Linux utility that can be used to report information about the status of all the hard drives (e.g. the number of bad sectors).

1. Research reported in this publication was most recently supported by the National Human Genome Research Institute of the National Institutes of Health under award number U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

There is a very large space of potential software solutions for these clusters. The goal of this abstract is to introduce readers to a core set of tools we find useful in developing and maintaining a low-cost cluster. In this poster, we discuss the tools we find most useful in efficiently managing our cluster and will provide a demo. We emphasize tools that are easy to learn and yet provide the necessary capabilities to manage a heterogeneous cluster. We provide support to a growing base of users on these issues.

REFERENCES

- [1] V. V. Kindratenko, J. J. Enos, G. Shi, M. T. Showerman, G. W. Arnold, J. E. Stone, J. C. Phillips, and W. m. Hwu, "GPU clusters for high-performance computing," *Proceedings of the IEEE International Conference on Cluster Computing and Workshops*, 2009, pp. 1–8. Retrieved from <https://doi.org/10.1109/CLUSTER.2009.5289128>.
- [2] K. Sajjapongse, T. Agarwal, and M. Becchi, "A flexible scheduling framework for heterogeneous CPU-GPU clusters," *Proceedings of the 21st International Conference on High Performance Computing (HiPC)*, pp. 1–11, 2014. Retrieved from <https://doi.org/10.1109/HiPC.2014.7116892>.
- [3] Lawrence Livermore National Laboratory, "Welcome to the Warewulf Project," 2017. Retrieved from <http://warewulf.lbl.gov/>.
- [4] The Redhat Corporation, "Ansible," 2017. Retrieved from <https://www.ansible.com/>.
- [5] Adaptive Computing, "The Torque Resource Manager," 2017. Retrieved from <http://www.adaptivecomputing.com/products/open-source/torque/>.
- [6] Adaptive Computing, "The Maui Job Scheduler," 2017. Retrieved from <http://www.adaptivecomputing.com/products/open-source/maui/>.
- [7] M. Massie, "The Ganglia Monitoring System," 2017. Retrieved from <http://ganglia.info/>.
- [8] C. T. Yang, T. T. Chen, and S. Y. Chen, "Implementation of Monitoring and Information Service Using Ganglia and NWS for Grid Resource Brokers," *Proceedings of the 2nd IEEE Asia-Pacific Service Computing Conference (APSCC 2007)*, 2007, pp. 356–363. Retrieved from <https://doi.org/10.1109/APSCC.2007.74>.
- [9] Y. Joshi, D. Sharma, U. Gaur, V. Kumar, and R. Kalmady, "Design of a Fault Tolerant Architecture for Private Cloud Computing Infrastructure," *Indian Journal of Science and Technology*, vol. 10, no. 4, 2017. Retrieved from <https://doi.org/10.17485/ijst/2017/v10i4/110663>.



Y+ * 2 / (0\$C1 - 1 . (9 (- /\$ = 2, - . \$ [0 ((\$1 -) \$X4 (- \$% * 0& (\$% "> ? 10 (\$R " " + 2

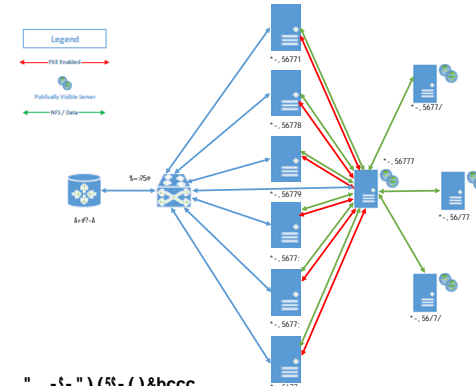
YB\$Y1 9 4 # (+ : \$S B C (& & 1 : \$ L B \$ X # (,) \$ 1 -) ! \$! B < , & " - (

R' (\$S (*01+ \$1 - , - . ((0, - . \$O1/1\$Y " - 2" 0/, * 9 : \$R (9 4 + (\$ = , 7 (02, /8

6#2/01&/

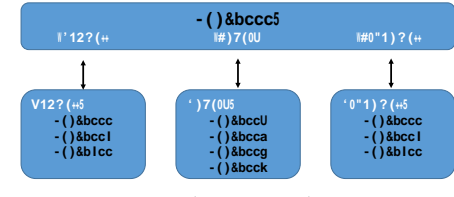
- L - \$"0 (0\$) \$ # (\$ > (& /, 7 : \$ 1 & " , - (\$ (10, - . \$ 9 * 2 \$ " 4 (01 / (\$ - \$ 40) # + (9 2 \$) \$ 2 & 1 + : \$ (0 (* , 0, - . \$ 2 * /, # 1 + 8 \$ + 10. (\$) 1 / \$ 1 -) \$ & " 9 4 / , - . \$ (2 " * 0 & (2 \$)

R' (\$S (*0" -, Q Y+ * 2 / (0



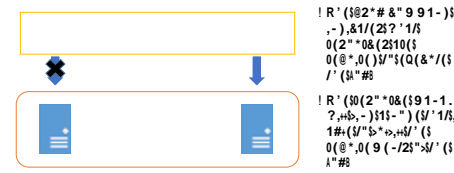
- ! kcc ; '\$%O: \$UKT\$; '\$N6C: UQ\$ C6C \$X4 / (0" - I \$Tamh\$ EU\$g\$; VZK
- ! Y < = \$ & " 9 4 * / (\$ -) (2 \$) () & bcc U \$ G (-) & dccc

Y+ * 2 / (0\$C1 - 1 . (9 (- /



- ! Y " 9 4 * / (\$ -) (2 \$ 1 0 (\$ # " / () \$ 0 " 9 \$ / (\$ - / ? " 0 P \$ E <] \ \$? , / ' \$ ^ 1 0 (? * > \$ E _ 1 ? 0 - & (\$ ' (0 P (+ \$ S \$ 1 , " - 1 \$ _ 1 # 2 \$

! "# \$ % * 9 , 2 2 , " -



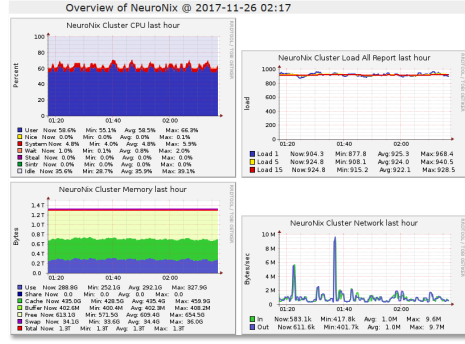
- ! "# \$ 2 1 0 (\$ 9 1 - 1 .) / \$ " 0 " . ' \$ / (\$ 0 * (* \$ 9 1 - 1 . (0 \$ 2 , - . \$ 1 \$ " 9 # , - 1 " - \$ " \$ R " 0 @ * (\$ E 0 (2 " * 0 & (\$ 9 1 - 1 . (0 \$ 1 -) \$ C 1 \$, \$ E 2 & () *) (0 \$)

! "# \$ % & ' () * + , - \$ % 0 / 1 (. , (2

- ! 3, 0 * 1 \$ 4 0 & (2 2 " 0 2 \$ 5 \$ 6 - \$ 1 # 2 / 0 1 / \$ * - , / \$ 4 0 (,)) # \$ 8 / ' (\$ 0 (2 " * 0 & (\$ 9 1 - 1 . (0 \$ 1 & 1 - # \$ (2 / \$) / \$ 0 (4 0 (2 (- / \$ & " 9 4 * / (\$ / 0 (1) 2 \$; < = 2 \$ " 0 \$ 1 \$. (- (0 1 \$ 9 (1 2 * 0 \$) > / ' (\$ 4 0 " & (2 2 , - \$ 4 " ? (0 \$ 0 (@ * , 0 () \$ " 0 \$ 1 \$ # \$

C - , / , 0 , -

- ! ; 1 - . , 1 \$ 9 " - / , 0 , - . \$ 2 8 2 / (9 \$ / 1 / \$ * + (& / 2) 1 / \$ 0 " 9 \$ -) (2 \$, - \$ 1 \$ & * 2 / (0 (0 , 0) \$ 1 -) \$, 2 4 1 8 2 / \$ (\$) 1 / 1 \$, - . 0 1 4 " , 1 & \$ " 0 9 \$ 0 " 9 \$ 1 \$? (# , - / (0 - 1 & \$



Y - > , * 0 1 / , - \$ C 1 - 1 . (9 (- /

- ! 6 - 2 , # , (\$ 1 \$ " " \$ / 1 / \$ 1 " ? 2 \$ " 0 \$ 1 " / 9 1 / , - / \$ (\$ 2 / \$ 4 5 1 - 0 2 " * 0 & (\$ 1 -) ; < = \$ " 9 4 * / (\$ -) (2 \$) 1 7 \$ / (\$ f . 4 " \$ 0 (2 " * 0 & (\$ 2 " # \$ 2 \$ 1 - \$ # (\$) , 0 (& /) \$ / \$ 0 " \$ - \$ - \$ (/ ' (0 \$ 1 Y < = 0 \$; < = \$ -) (\$)

< 0 " 7 , 2 , " - , - . \$? , / ' \$ ^ 1 0 (? * >

- ! L - \$ (1) , / , - \$ / " \$ 4 (0 1 , - . \$ 2 8 2 / (9 \$, 9 1 . (2 \$ 1 - 8 \$, > (\$ 0 " 9 \$ / ' (\$ 9 1 2 / (0 \$ -)) (\$ 1 - \$ # (\$ 9 1) (\$ 1 7 1 , + # , \$ / \$ / ' (\$) & " 9 4 * / (\$ -) (2 \$)

% 9 9 1 0 8

- ! 6 \$ / (/ 0 " - (- (* 2 \$ & * 2 / (0 \$ 2 , - . \$ 1 \$ 0 ((\$ 2) > / ? 1 0 (\$ 2 / 1 & P \$ 0 * - - , - . \$ - \$ 1 \$ & " + & / , " - \$ > , - (0 4 - 2 , 7 \$ & " 9 4 * / (\$ -) (2 \$) & 1 - \$ # (\$ 2) / \$ / \$ " ? 0 P \$ " (\$ 7 - / \$) (\$ 9 2 / \$ & " 9 4 * / 1 / , - 1 + 8 \$, - / (- 2 , 7 \$ 4 0 # + (9 2 \$)