How To Design a Cluster

PRESENTED BY

ROBERT C. JACKSON, MSEE

FACULTY AND RESEARCH SUPPORT MANAGER

INFORMATION TECHNOLOGY STUDENT ACADEMIC SERVICES GROUP

THE UNIVERSITY OF TEXAS-RIO GRANDE VALLEY

How To Design a Cluster

***Abstract*—What has evolved to be known as Cluster Design, has its roots in Cluster Building. Cluster Building is essentially the procurement of Commodity Off The Shelf (COTS) components followed by the 'brute force' assembly of said parts, including PC's, data switches and storage units into a connected group of systems with the purpose of achieving advanced collective computational capability or availability. An Operating System (OS), homebrew scripts, ssh keys, and a scheduling mechanism completed the build, after benchmarking the cluster was ready for general usage. Today's Cluster Design process that supplanted Cluster Building is based on the types of problems the system will need to solve, optimizing performance and future expansion. Evolution and advancement of COTS components and HPC applications accommodate different levels of operation and performance making design and fitting of components essential for assembling Clusters that meet the needs of and provide optimum performance for user and systems applications.**

How To Design a Cluster

Introduction

Before deciding to build or purchase a High Performance Compute (HPC) Cluster the problems that it will be used to solve must be taken into consideration. This forethought will decide how various hardware and software options can be interleaved into a total cluster fabric. Performance expectations above that of PC's and Servers, utilization and resource consumption must also be taken into account. Finally, how this resource will be managed and operated, and cost must be tallied. All of this input evaluated effectively will help decide the ultimate design of a cluster, whose operation and usage will be finely tuned to the target audience of users.

How To Design a Cluster

BACKGROUND

A. The good old days...

What is a Cluster? A Cluster is a bunch of computational resources (PC's) connected together to work on problems where combined each computer works on their part, with the results collected at some common point of interconnection between all of the systems.

Conceptually Engineers, Scientists and Technicians envisioned advanced computing capabilities of high performance CPU's, memory and storage (Von Neumann Model) as far back as the 1960's. Complex Instruction Set Computing (CISC) the processing methodology of the day whose large instruction set which is like a high level language like 'C, i.e. a lot of instructions, e.g. VAX. Reduced Instruction Set (RISC) CPU's have a simpler instruction set which is very adaptable to upscaling (increased clock speeds, pipelining, cache coherency, Out-Of-Order Execution and Branch Prediction). MIPS and Sparc are examples of RISC CPU's. RISC CPU's haven't been very successful in the commodity and Desktop PC market dominated by X86 architectures.

Note: current clusters mainly employ x86_64 Intel CISC CPU's.

How To Design a Cluster

Background Part A cont'd

RISC/CISC CPU Equation

The Performance Equation

The following equation is commonly used for expressing a computer's performance ability:

$$\frac{\text{time}}{\text{program}} = \frac{\text{time}}{\text{cycle}} \times \frac{\text{cycles}}{\text{instruction}} \times \frac{\text{instructions}}{\text{program}}$$

The CISC approach attempts to minimize the number of instructions per program, sacrificing the number of cycles per instruction. RISC does the opposite, reducing the cycles per instruction at the cost of the number of instructions per program.

How To Design a Cluster

Background Part A cont'd

Memory, Storage and Connectivity

Factoring in memory, storage and connectivity the 'cluster of old' was very capable. Main memory (DRAM) had a lesser architectural footprint compared to the modern systems of today. CPU Cache memory was smaller too. Storage capability wasn't as dense with PATA drive capacities in the GB range and connectivity was typically 10Mb/s or 100Mb/s Ethernet.

How To Design a Cluster

Background Part A Con't

From this technology the Network of Workstation (NOW) type cluster came into being. This basic cluster was pieced together with older sometimes throw away parts – an Ethernet switch and IP cables, several PC's probably destined for surplus, an OS customized with scripts and compilable source code, benchmarked to some MFLOP value. Beowulf clusters were the successor to the NOW cluster and were succeeded by what we have today.

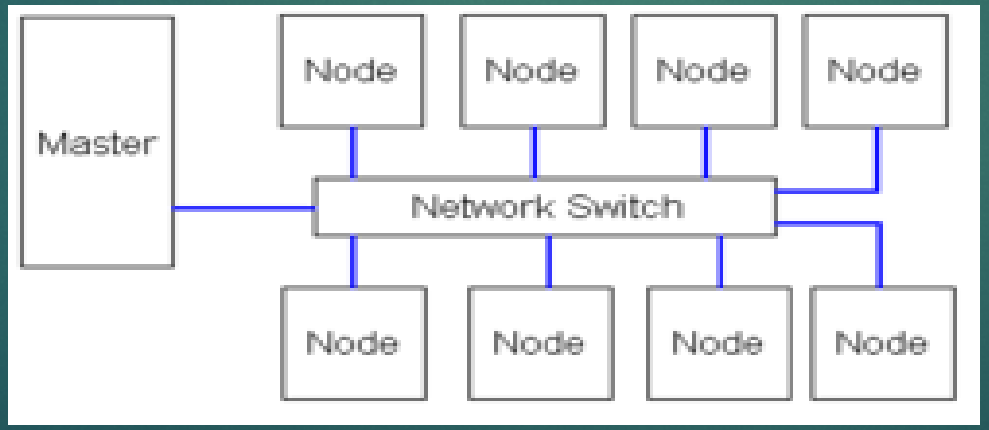

Figure 1 Cluster Block Diagram

How To Design a Cluster

Background Part A cont'd

*W*ith time the basic Cluster Build gave way to a more robust and *specific* Cluster Design which takes into account many specifiers, including multicores, compiler optimization, fast I/O, advanced middleware and resource management, large storage capability, fault tolerance and power consumption. Clusters have come a long way from connecting a few old PC's together with a switch, generating some ssh keys, writing some shell scripts for operability and compiling some 'C' code into some sort of batch scheduler.

How To Design a Cluster

Background

B. Today's Clusters

The cluster of today is infinitely more complex than those early clusters and more powerful by magnitudes of millions, billions or trillions. The MFLOP rating of old has been replaced by TFLOP and PFLOP ratings. Nodes with 20 or more CPU cores are commonplace, as are nodes with up to 1TB of main memory for memory intensive applications. High speed I/O and filesystems are standard issue, and many systems have co-processors and GPU cards to mediate the computational load. Large Storage devices provide Terabytes and petabytes of space for use as scratch space to work problems or for back-end storage. Work and resource managers are of intelligent design and run right out of the box with default settings-but can be customized for ease of use. Compilers and system libraries can be selected for easy installation and accommodation of the multitude of applications that run on today's clusters.

How To Design a Cluster

Background

C. Clusters at UTRGV

Current cluster technology at the University of Texas-Rio Grande Valley includes 3-clusters: 2-IBM, 1-DELL, all older but still very capable systems. About 10% -15% of our faculty and researchers use HPC resources for research and/or teaching, this is an area of improvement for us.

The IBM cluster Thumper is used for general MPI, OpenMP, matlab, Stata, Schrodinger, RSoft, NWCHEM, NAMD and Velvet jobs. Using the Grid Engine Scheduler this cluster has several queues and parallel environments to handle the aforementioned applications. A recently added GPU node with 2-nVidia GPU's handles cuda jobs.

How To Design a Cluster

Background

C. Clusters at UTRGV Cont'd

Thumper Cluster Architecture

Thumper Specs	Description
NW Media	1-GB Ethernet (public, private), FDR-Infiniband
Node Specs	1-Mgmt, 2-Login – IBM x3550 m3, 1-Storage – IBM x3650 m3, 68-Compute – dx360 m3, 1-GPU – x3650 m5
Cores	Mgmt, Login, Compute – 2-dual socket, 6-core; Storage – 2-dual socket, 12-core; GPU – 2-dual socket, 20-core; 916 cores total
Memory	Mgmt, Login – 12TB; Compute – 24TB; Storage – 48TB; GPU- 28TB; 1732TB total memory
Storage Array	DS3512 – 24TB, ext4, NFS
OS	RHEL v6.7
Resource Mgr.	Grid Engine
Cluster Manager	xCAT

How To Design a Cluster

Background

C. Clusters at UTRGV Cont'd

Thumper Cluster

How To Design a Cluster

Background

C. Clusters at UTRGV Cont'd

The Futuro Cluster which is 7 years old and currently under maintenance is used for numerical methods simulations and is comprised of components listed in following table.

Futuro Specs	Description
Nodes	1-MGMT, 2-Login, 1-Backup, 2-Storage, 4-Student, 30-Compute
NW media	1-GB Ethernet (internal), 1-GB Ethernet (public), FC-Infiniband
Node Specs	Mgmt, Storage, Backup – IBM 3650 m2; Login, Student, Compute – IBM dx360 m2
Cores	All – Dual socket Quad core: 320 total
Memory	Mgmt, Login, Compute – 48GB; Storage, Backup, Student – 24GB
Storage Array	Storage – 192TB, Backup Storage – 24TB; GPFS filesystem.
OS	RHEL v7.1
Resource Mgr.	Slurm
Cluster Mgr.	xCAT

How To Design a Cluster

Background

C. Clusters at UTRGV Cont'd

Our oldest cluster Bambi is an 11-year old DELL cluster and is currently used for teaching and as a Linux Lab. Bambi is comprised of components listed in the table.

Bambi Specs	Description
NW media	1-GB Ethernet (private, public)
Node Specs	1-Head, 1-Storage – Dell PE 2950; 1-Mgmt, 55-Compute – DELL PE 1950
Cores	Mgmt, Storage, Head – 2-dual core, 12 cores total; Compute – 2 dual core , 220 cores; 232 cores total
Memory	Mgmt, Storage, Head, Compute – 4GB; 232 GB total.
Storage Array	2TB ext4, NFS
OS	Centos v6.6
Resource Mgr.	Grid Engine
Cluster Mgr.	RocksClusters v6.2

How To Design a Cluster

Background

C. Clusters at UTRGV Cont'd

Bambi Cluster

How To Design a Cluster

Cluster Design

Moving Forwards UTRGV needs a cluster that can span and scale across its wide campus area consisting of two large campuses 50 miles apart, a medical school and 2 smaller campuses also spread across a large area.

A. Now the future…

Taking into consideration our current cluster architecture, customer base and usage, matlab, NAMD, Schrodinger Suite and Stata are our most run applications. Our GPU utilization is very low as are SMP and MPI job runs. These are additional areas of improvement. Taking this and funding into account a high speed parallel file system like Lustre or GPFS filesystems may not be needed, high speed NFS may suffice. Infiniband and/or OmniPath interconnects have become standard on the latest Clusters, but high speed 10G Ethernet is also a player and is cost effective.

How To Design a Cluster

Cluster Design

B. Components

Systems with multiple cores and large memory, can solve problems specifically for jobs that require SMP and need many cores. Large memory systems can accommodate jobs that do a lot of calculations and are memory intensive. Matlab jobs could also benefit from additional CPU cores and large memory on compute nodes.

- CPU selection must be leveraged against price/performance and node quantity.

- Intel E5-2680v4 CPU's provide good computational capability, allowing for more compute nodes then a more powerful processor.

- All nodes will have between 64TB -256TB of memory.

How To Design a Cluster

Cluster Design

B. Components

GPU and/or Xeon Phi nodes can mediate the processing of cluster jobs affecting load and increasing the execution performance reducing execution time. Bearing in mind there's a learning curve for both solutions, a premium price to performance ratio and that a lot of users don't have applications that make use of this technology, a small investment here would make sense. Deep Learning and Machine learning projects would benefit from these technologies.

Selecting a Co-Processor or GPU configuration will depend on the applications that will be run and researchers/users experience level.

- NVidia GPU's are preferred because of their association with Deep Learning and Machine learning.

How To Design a Cluster

Cluster Design

B. Components

InfiniBand, Omni-Path or 10Gb/s Ethernet are all effective connectivity standards that can be leveraged against funding and application requirements. InfiniBand (IB) is expensive but offers high throughput and very low latency working best with I/O based applications. Omni-Path (OPA) a newer technology and direct competitor to IB offers competitive low latency/High throughput figures as those of IB at lower cost. 10Gb/s Ethernet still has the higher latency due to TCP/IP overhead but offers greater throughput due to faster speed and is an economical connectivity media.

- 10Gb/s Ethernet is a good choice for connectivity based purely on price and funding.

How To Design a Cluster

Cluster Design

B. Components

Fault tolerance may necessitate redundancy being designed into the cluster e.g. dual login nodes, dual mirrored storage arrays, 2 or more management nodes, extra switches/switch ports. This way in the event of failure the system can fail over to these duplicates. System upgrades can also be made with minimal effect on normal operation.

- A cluster with dual management nodes, dual login nodes, dual storage arrays, a single storage system and fitted switches would suffice to make the cluster sufficiently fault tolerant.

How To Design a Cluster

Cluster Design

B. Components

Large storage is ultimately necessary to allow users to run their apps storing intermediate run files and final result files, provide common space between nodes and be a scratch area for testing submissions.

A Backup capability is necessary to archive user data to prevent data loss, by accidental deletion corruption or otherwise. In some cases data backup is the responsibility of the data owner and not a systems operation function.

GPFS or Lustre the big 2 distributed filesystems (dfs) are almost a no-brainer for the current HPC Cluster. GPFS is a mature, subscription-based dfs from IBM, Lustre is open source, faster than GPFS but doesn't work well with small files (I'm told). There's also Ceph dfs and Gluster dfs. High Speed NFS is also an option to be considered where funding is concerned.

How To Design a Cluster

Cluster Design

C. The Ultimate cluster

A cluster for UTRGV would have at least 150 nodes, 40 of which can be configured for teaching and training use, the rest configured for dedicated research usage and would cost ~$930K. The cluster would look like this:

- 140 compute nodes inc/4 GPU nodes w/4 nVidia GPU's per node, and a few large memory nodes
- 2 – management nodes
- 2 – login nodes
- 2 – storage arrays: 500TB each (mirror the most active filesystems)
- 1 – storage system
- 3 – GPFS systems or use NFS (these 3-nodes become execution nodes)
- 3 – 10 Gb/s Ethernet switches
- RHEL v7
- Bright Cluster Manager
- Slurm Resource Manager

How To Design a Cluster

Cluster Design

D. Other things to consider

- Training
- Overfitting/underfitting

Fitting Factor 1	Fitting Factor 2	Fitting Factor 3	Fitting Type
Utilization Saturated	Arch. < Adeq.	Str. Saturated	Underfit
Utilization average	Arch. = Adeq.	Storage = Adeq.	Fit
Utilization low	Arch. > Adeq.	Storage >>> Adeq.	Overfit

- Warranty/service – 3years or more?
- funding for miscellaneous expenditures
- Power consumption
- Citations, publications, conference presentations

How To Design a Cluster

Q&A