Big-FE - A cost and power efficient platform for high performance computing education

Alex Lemann(*), Kevin Hunter(*), Joshua McCoy(*), Charles Peck

Earlham College, Richmond, IN

Abstract

There is a shortage of high performance computing (HPC) resources for educational use. Big-FE, short for Focus on Education, is a complete design and hardware prototype for a cost effective, energy efficient, and scalable HPC resource. Our design uses a diskless compute node architecture running the open source Bootable Cluster CD distribution. Big-FE is a synergy of the classic Beowulf commodity-off-the-shelf cluster philosophy and the commercial blade approach to computational cluster design. Leveraging the strengths of each approach, Big-FE provides a cost effective computing platform for computational science education.

We measure the energy consumption of the Big-FE prototype and compare it with popular commercial HPC platforms. Our physical layout is organized to minimize both maintenance and cooling costs. We detail the full cost structure and estimated performance for Big-FE and compare these values with those of popular commercial HPC platforms.

Keywords: education, hardware

Motivation

Computational Science Education

Recent advances in science rely on high performance computing (HPC) platforms and computational methods, yet the calculation of π is still the canonical example motivating use of parallel computing techniques. Examples like this lack the relevance and luster necessary to connect theory, real-world science, and most of all learning. How can we teach the nuances of parallel matrix-matrix multiplication to undergraduate students who have at best only solved linear systems with their computational mathematics package (e.g. Octave, Maple, or Mathematica) or at worst, have never been confronted with the need to solve a linear system that could not be solved in an instant using the computational capabilities of today's handheld devices. This fundamental disconnect first affects our ability to attract students to computational science, and later our ability to adequately keep and prepare students for graduate school or professional work in HPC and computational science.

The lack of access to High Performance Computational resources in the classroom makes it particularly difficult to effectively teach students about problem decomposition, speedup, and efficiency– key aspects of using an HPC resource. Big-FE seeks to address this by providing a cost effective alternative to traditional HPC platforms.

Green Computing

Power and cooling have become the principle design challenges for cluster engineers at all scales. As an example, if one uses a cost \$0.15 per KWh for the cost of electricity, the typical rack full of 32 name-brand 1U dual processor servers will cost more to power and cool over 4 years than it cost to purchase new.

Big-FE is engineered from the ground up to minimize the amount of power required and therefore the amount of heat that needs to be dissipated.

What is Big-FE?

Big-FE is a blade architecture, Beowulf style, high performance computing (HPC) cluster. Big-FE's design is a synergy of the classic Beowulf commodity-off-the-shelf cluster philosophy and the commercial, proprietary blade approach to computational cluster design. Big-FE leverages the strengths of each approach to provide a cost effective computing platform for computational science education.

Design

Big-FE's design was guided by our experience with building and operating multiple Beowulf style computational clusters over the past 5 years: Athena, bazaar, Cairo, sandbox, ACL, UNI, and Little-Fe, among others. We were also guided by the Beowulf list archives, Brown's "Engineering a Beowulf-Style Cluster", Google's experience, and particular blade manufacturers.

Our goal was to design a system that met the needs of teaching colleges with limited administrative resources, a system that would be power and cooling conscious and engineered for maintainability. The system is designed with a central chimney, allowing airflow where cooler air can be blown in from the bottom of the rack and rise through the top of the rack. The use of sub-racks combined with the open design allows for easy assembly and maintenance.

A Big-FE cluster consists of one or more standard 19 inch equipment racks, each of which is populated with a standard set of mounting and computational components. Each Big-FE rack unit consists of 8 sub-racks, mounted 4 to a face, each of which holds 8 dual processor blades, for a density of 128 CPUs per rack unit. Blades within each cluster are connected to local gigabit Ethernet switches. Each Big-FE rack is connected to one or more co-located Big-FE racks with gigabit multi-mode fiber interconnects.

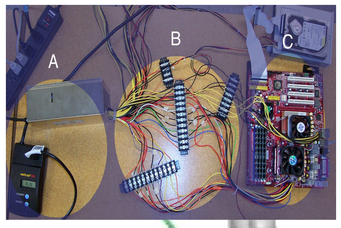
Each blade is an MSI K8T Master2-FAR7 dual processor motherboard outfitted with 2 AMD Opteron (64 bit) 246 CPUs. The MSI motherboards are dual core capable providing Big-FE with an efficient upgrade path. In the x86 world dual processor motherboards are currently at the sweet spot of the density-cost continuum. AMD's Opteron CPUs have proven themselves to be cost-effective and power-efficient when compared to similar x86 offerings. The MSI motherboard offers a number of important features: a Broadcom gigabit network interface, support for LM sensors, and a BIOS which supports disabling of unused components (e.g. SATA interface, video) which will enable us to reduce power consumption.

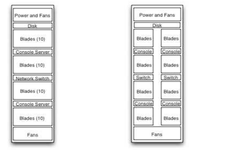
All of the components used in our design are commodity-off-the-shelf and can be sourced from multiple vendors for competitive bidding.

Software

BCCD

The Bootable Cluster CD (BCCD) has been developed by Paul Gray as a Linux distribution focused on being able to turn any lab into a cluster through a bootable CD. In this form it does not modify the underlying system. The system is configured with various curriculum modules which explore topics in computational science as well as standard clustering tools. These can be downloaded through the system's list-packages software management tool. Big-FE plans to leverage the work already in the BCCD by using a "liberated" form of the BCCD, one which has been installed on a hard-drive.





Single Big-FE rack unit (typical). Each rack unit is packaged in a standard 77" tall 4 post EIA/TIA rack (44U tall x 19" wide x 30" deep). Each institution would house 2 rack units.

Measuring Power Consumption

In order to determine the characteristics of a power supply capable of powering one of our sub-racks, we needed to measure detailed information on the power consumption of a single motherboard/blade. The total DC amperage per wire on each of the standard ATX connectors comprise a single node power consumption profile. Additionally, a separate power consumption profile for different load conditions are used to determine the maximum power output capacities the power supply would need.

The total AC amperage is measured by a WattsUp power meter set inline between the wall jack and the power supply (noted as "A" in the figure). Every wire in each connector that powers the motherboard was cut and reconnected to a terminal block ("B" in the figure). A multi-meter was then inserted to measure the DC amperage used by each voltage rail.

The sum of the DC amperages used on each rail, with an additional safety factor, determine the maximum output capacity needed on that rail of the power supply for a single node. The table below shows the amperage used on each power supply rail with varying load profiles.

Rail	Idle	CPU	Disk	2CPU
3.3V	5.89	5.91	6.62	7.13
5.0V	.59	.59	.86	.59
12.0V	6.35	7.41	6.76	8.19

Multiplying the single motherboard/blade values by a factor of eight results in the necessary values for a power supply needed to power a single subrace.

Commercial Solutions

Currently there are two common architectures for HPC clusters: the standard rack (1U or 2U) servers and blade servers. Blades offer a higher density of computers per rack as well as a higher power supply efficiency. Typically Beowulf clusters – those built from commodity-off-the-shelf hardware – are limited to 1U/2U servers because there are no standard power supplies built for multiple motherboards. Big-FE's design offers an expansion of the standard Beowulf cluster design into the domain of blade servers with the benefits of density and efficiency. In Figure 1 and the subsequent figures, "Unit" refers to either a single 1U server, a single blade, or a single motherboard blade as in Big-FE.

A standard purchase procedure would be to look to commercial vendors for ready made cluster solutions. Relying on a commercial vendor provides stability because there will be a support network in place. Also, commercial vendors typically have well tested and integrated hardware and software solutions.

On the other hand, the commercial route also introduces several potential problems. First, a corporate vendor has spent a significant amount of time and money designing and testing their HPC solution – an investment that will need to be recouped on every order to their company. Second, leaving the "heavy work" to an HPC vendor removes the potential for students to be involved in the design and testing process. Lastly, HPC vendors are tied to particular suppliers for their hardware and software. This means that the customer will consequently be locked in to their vendor's choices.

Choosing the Big-FE HPC model allows the customer to be involved in each aspect of the design, but does not lose all of the simplicity afforded by purchasing a ready-made solution from an HPC vendor. While a Big-FE HPC system is not pre-assembled, the plans and instructions give a general design framework from which to work. This means that smaller institutions do not need to strain their budgets purchasing access to an HPC environment. They do not need to spend more money during the planning phase than with a vendor supplied solution and they gain the benefits of a cost effective and powerful HPC cluster.

Comparing Big-FE's initial purchase cost and annual upkeep to a similarly powerful 1U Dell cluster and a IBM BladeCenter cluster show the efficiency and density that can be gained from a custom blade server design. Big-FE offers significant savings with an initial purchase cost of \$95 dollars per Gigaflop compared to the \$178 dollars per Gigaflop of a Dell 1U SC1425 based cluster and the \$485 per Gigaflop for an IBM LS20 BladeCenter cluster. Big-FE offers further savings in the yearly maintence cost of \$10 dollars per Gigaflop. Ten dollars compares nicely to the LS20's \$14 per Gigaflop annually and the SC1425's \$20 per Gigaflop annually. Here the efficiency of the DIY blade design is apparent, but the annual savings is given at the expense of the high development costs for blade clusters compared to one created with standard 1U computers. In either case, the commodity-off-the-shelf approach of Big-FE presents a significant cost savings over pre-packaged commercial systems.

Cluster Vendor	Units per Rack	CPUs per Unit	CPUs per Rack	GFlops per CPU	GFlops per Rack
Dell 1U SC1425	32	2	64	6.4	409.6
IBM LS20 Blade	56	2	112	4.0	448.0
Big-FE	64	2	128	4.0	512.0

Cluster Vendor	Cost per Unit	Cost per Rack
Dell 1U SC1425	\$2,282	\$73,014
IBM LS20 Blade	\$3,882	\$217,404
Big-FE	\$758	\$48,480

Cluster Vendor	KW per Unit	KW per Rack
Dell 1U SC1425	0.3310	10.59
IBM LS20 Blade	0.1429	8.00
Big-FE	0.1080	6.91

Cluster Vendor	Power Cost per Rack (Annual)	Cooling Cost per Rack (Annual)	Total Operating Cost per Rack (Annual)
Dell 1U SC1425	\$6,495.01	\$1,855.72	\$8,350.73
IBM LS20 Blade	\$4,905.70	\$1,401.63	\$6,307.33
Big-FE	\$4,238.44	\$1,210.98	\$5,449.42

Future Work

Our next task is to engineer a single power supply that will feed all 8 blades in a single sub-rack. Larger power supplies are more efficient, and the reduced count will make mounting and cooling easier. The power supply calculations which we performed (see the chart above) will form the basis for this design.

The second task will be the completion and testing of a single sub-rack prototype. This will enable us to work-out the wire management and airflow details of the design.