CLEARSPEED PAPER: A STUDY OF ACCELERATOR OPTIONS FOR MAXIMIZING CLUSTER PERFORMANCE

Abstract

The typical primary goal in cluster design now is to maximize performance within the constraints of *space* and *power*. This contrasts with the primary goal of a decade ago, when the constraint was the budget for parts purchase. We therefore examine how to create a cluster with the highest possible performance using current technology. Specific choices include a pure x86 design versus one with accelerators from ClearSpeed, or even future 64-bit versions of NVIDIA's announced "Tesla" product. We use compute-per-volume arguments to show that a cluster achieves optimum 64-bit floating-point performance with x86 nodes enhanced with ClearSpeed e620 "Advance" accelerators. The reason is that power dissipation and its attendant volume become the primary limiters to maximizing cluster performance.

The Constraints of Modern Cluster Design

There are many constraints and goals in cluster design, and we do not claim to address all of these here. In this paper, we are concerned chiefly with physical limitations: electrical power, heat removal, floor space, floor loading, and volume.

We recognize that many other issues enter into the decision process. For example, we implicitly assume that this is a cluster that uses mainstream technology such as x86 processors and the Linux operating system (as opposed, say, to proprietary designs such as Blue Gene, Cell BE, or GRAPE). We assume that the cluster must be fast and accurate at 64-bit floating-point operations (full precision) for HPC applications. We also assume that any high-capability cluster must have resilience to errors, which means that it must be able to detect and act on errors instead of silently allowing them; it also means that it must be engineered with sufficiently reliable components that its Mean Time Between Failures (MTBF) will not be intolerably small.

The Coupling between Volume and Power

Perhaps the most important constraint as of mid-2007 is the following: **Air-cooled computers can dissipate a maximum of about 70 watts per liter.** If a particular component exceeds this power density, it forces the addition of unproductive space elsewhere, both for venting of air and for additional power supplies beyond what standard platforms are designed to provide. Beyond 70 watts per liter, temperatures rise above operating limits, even with heroic efforts to engineer the airflow.

It is interesting to test this guideline on the standard PCI slot specifications. The following table shows the volume in liters and corresponding wattage (at 70 watts per liter) for various standard PCI form factors:

Table 1. Implied wattages for PCI geometry

Form factor	Volume, liters	Watts, at 70 W / liter
Full-length, full-height	0.676	47.3
Half-length, full-height	0.363	25.4
Full-length, low-profile	0.406	28.4
Half-length, low-profile	0.216	15.3

The standard power limit on PCI-X and PCIe slots is 25 watts, which is in line with the rightmost column in Table 1. Some boards combine power from multiple rails to exceed the standard, which potentially increases the effective volume demanded by the board beyond its geometry.

We can also test the 70 watts per liter guideline at the server level. The current 1U (1.75 inches) servers consume about 1000 watts maximum, including extensibility options. The standard width is 19 inches, and a typical depth is 26.5 inches. The volume of the 1U server is thus about 14.4 liters. At 70 watts per liter, this volume allows almost exactly 1000 watts.

At the rack level, we empirically observe that air-cooled clusters hit a limit of about 40 kilowatts per rack (42U high rack). Note that clusters usually need about 6U of the rack for communication, uninterruptible power supplies, additional cooling, and service processors. This limits the maximum number of 1U compute servers per rack to about 36. Blade servers offer some advantages, but the limits to power density apply to them as well and thus the arguments here apply regardless of the vertical or horizontal orientation of the nodes. The volume of the standard rack is about 600 liters, which corresponds to a maximum of 42 kilowatts by the 70 watts per liter guideline.

Thus we see ample empirical evidence that high-performance air-cooled clusters cannot exceed 70 watts per liter.

Facility Constraints: Power, Space, and Weight

The electrical power available for a proposed cluster is always constrained. Depending on the facility where the cluster is located, typical limits range from one to eight megawatts. The arguments here scale easily within this range, so we will illustrate with a two-megawatt limit on total power. While this might seem to imply we can have 50 racks consuming 40 kilowatts each, this is not the case. Much of the power must go to *removing* the heat generated by the racks. The fraction varies, but a typical guideline is that 40% of total power must go to cooling. Thus, only 1.2 megawatts are available for the computing nodes themselves and their communication fabric. If each rack uses its entire 40-kilowatt budget, this implies 30 racks.

Total floor space can also be limiting. When we include space for egress and cabling and airflow, a good rule of thumb is ten square feet (0.93 m^2) per rack. A cluster with 30 racks would then take up 300 square feet. To this we must add space for the chillers, uninterruptible power supplies, disk and tape storage, and so on; these are not performance-limiting parts of the system like the computing and communication equipment because they can be placed some distance away.

Finally, *weight* can be a limit, because the strongest racks cannot hold more than 1200 kg total, and

400–600 kg is a more typical maximum capacity, in addition to the roughly 100 kg for the rack itself. The floor load limit for a raised floor computing center is usually 250 pounds per square foot, or 1222 kg/m^2. Since the exterior enclosure of the rack is effectively 24 inches by 31.5 inches, its footprint is 0.49 m^2 and thus the weight cannot exceed 1222 × 0.49 = 596 kg. If we allow 96 kg for the rack itself and the cabling, the 1U slots cannot hold more than 11.9 kg (26 lb) each, on the average.

Communication Constraints: Interprocessor and Memory Bandwidth

Whatever communication fabric one selects (Ethernet, InfiniBand, Myrinet, etc.), one must add high-performance switches. The best way to combine switches with a higher-level switch is beyond the scope of this paper, but in general, it is possible to couple 30–40 racks with a two-level hierarchy using commercially available parts. If the intended application demands full performance for arbitrary interprocessor communication patterns, then switches can take up much of the space and power. We estimate that 10% of the rack space and power suffice for an effective high-performance cluster. This leaves 0.9 × 1.2 MW = 1.08 MW for the computational nodes themselves. That translates to 15,400 liters of computing hardware after accounting for all the overheads.

An equally important constraint is the bandwidth of the x86. Current "Bensley" platforms supply 19 GB/sec between the DRAM and the two x86 sockets, and the next generation will soon provide 24 GB/sec. This restricts the number of PCIe interfaces the platform can reasonably support, to about eight PCIe 8-lane ports.

The next section will consider what types of computational "bricks" we can build to fit into these constraints.

Building Block Densities

As an example of the argument, we can use two computing solutions that are readily available as of mid-2007: a standard x86 server and the ClearSpeed e620 "Advance" board. We also can consider a future 64-bit version of NVIDIA's "Tesla" product, which NVIDIA claims will be available in late 2007. While NVIDIA has not formally announced what the 64-bit performance will be, we can use estimates provided by their company representatives in public presentations.

Figure 1 shows three volumes, drawn to scale, for the current ClearSpeed Advance board, the future 64-bit version of the NVIDIA Tesla board set (where we assume the same form factor as the present C870 board set), and a standard 1U server.

The 1U server could be an x86 server with the highest compute density (presently a dual-socket "Clovertown" Intel Xeon 5300 running at 2.66 GHz). The total volume in the 1U server is 14 liters. Approximately 10 liters is required for the base processing hardware, memory, power, communications, and so on. As Figure 1 shows, about 4 liters of volume (about 300 watts) is typically available for expansion options, which could be local disk storage, additional DRAM, or PCI slots. Note that while eight ClearSpeed boards fit in that space and power budget, only one of the future Tesla board sets will fit.

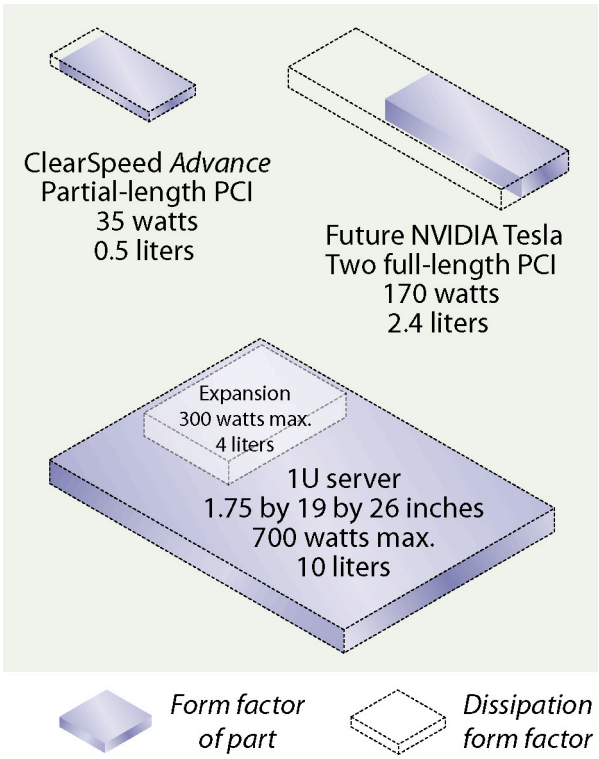

Figure 1. Compute Densities

We could consider the NVIDIA packaging of four of their Tesla board sets in that same 1U volume, a 64-bit version of their announced "S870" product, and again assume their 64-bit product has the same form factor. However, the NVIDIA S870 still requires a host server, so the NVIDIA S870 option actually demands 2U of rack space.

Density of GFLOPS and GB

The ClearSpeed Advance board has a peak 64-bit speed of 80.64 GFLOPS in both its PCI-X and PCIe versions, and presently ships with 1 GB of DRAM memory.

The announced NVIDIA Tesla product line creates an apples-to-oranges comparison unless we exercise care. NVIDIA's GFLOPS claims are for 32-bit floating-point only, and even that half-precision is barely suitable for many HPC applications because the results are not correctly rounded to the nearest number (GPUs truncate bits, so iterative

computations decay toward zero). Many intrinsic functions are only accurate to within five Units in the Last Place (ULPs), which is an order of magnitude less accurate than typical HPC intrinsic functions. Furthermore, NVIDIA counts *three* floating point operations per cycle; the multiply-add can be supplanted with an additional multiply in another functional unit, but it is almost universal in HPC to require and measure peak FLOPS with a one-to-one ratio of multiplies and adds. The maximum rate at which a Tesla GPU can create half-precision truncated multiply-adds is 345.6 GFLOPS.

The present NVIDIA C870 board set has no 64-bit floating-point hardware. NVIDIA stated in public presentations in June 2007 that the 64-bit performance of the future version of Tesla, which we annotate here with a "+" after their product names, would be *one-eighth* that of its 32-bit performance. For HPC purposes, we use the multiply-add speed and ignore the separate multiplier unit; this means the present C870 product has a peak 32-bit speed of 345.6 GFLOPS. If the C870+ has the same 32-bit speed, this means the 64-bit performance will be 43.2 GFLOPS. If NVIDIA is able to, say, double their 32-bit speed, then the 64-bit speed could be as high as 86.4 GFLOPS.

The C870 board set has 1.5 GB of DRAM. We assume this will not change for the C870+.

For the future 64-bit version of their 1U "S870" product, multiply all numbers by four: 6 GB of DRAM and an estimated full-precision speed of 172.8 – 345.6 GFLOPS. There have been suggestions of Tesla boards with two GPUs instead of one, but this would double power consumption to 1600 W, more than a 1U space can dissipate. Assume the Tesla S870+ will be 172.8 to 345.6 GFLOPS peak, not counting the additional 1U host server that drives it.

A 1U x86 server draws 1000W maximum (including expansion options). With two sockets (four cores per socket), the current Xeon 5300 has a peak 64-bit floating-point speed of 2.66 GHz times 32 floating-point operations per cycle = 85.12 GFLOPS. Memory densities currently permit 32 to 64 GB in a 1U space, depending on the amount reserved for PCI expansion. For maximum computing power, we assume 32 GB of DRAM in the server, and that all 4 liters of expansion volume can be dedicated to PCI slots.

Table 2 summarizes the computing and memory capability per unit volume. Note that the NVIDIA boards significantly dilute the memory per liter, whereas the ClearSpeed boards maintain a memory density comparable to that of the x86 server.

Table 2. GFLOPS and GB per liter

	GFLOPS per liter	GB per liter
ClearSpeed e620 board	161	2
NVIDIA C870+ boards	18–36	0.6
NVIDIA S870+ 1U (no host)	12–24	0.4
Dual Intel 5300 server	6	3.2

Weight and floor loading

We cannot create a table for weight in kg of each option because there is no published data for the Tesla product, and even the 32-bit version of the product is not yet shipping at the time of writing. However, we note that a ClearSpeed e620 board weighs only 250 g, so even if we were to pack 8 of them into a 1U server, they would only add 2 kg to the 1U enclosure. The Tesla products include fans and power supplies, so we can expect them to be both denser and heavier.

A simple rule of thumb is that the weight in pounds of a 1U device should not exceed its rack depth in inches. If it does, it contributes more than its share of the 250 lb/ft^2 load limit. We can readily find 1U dual-Xeon 5300 servers that weigh over 35 pounds, such as the Dell PowerEdge 1950. However, this server is more than 26.5" deep, and hence it exceeds the load limit by only about 8% if it populates every 1U slot.

Summary of facilities constraints

In summary, every facilities aspect of cluster design seems to be hitting its maximum: available power, watts per liter, and kilograms per square foot. If any "building block" in the system exceeds limits, it simply forces the use of nonproductive empty space elsewhere in the system. While the future NVIDIA Tesla 64-bit product may offer a 2x to 4x compute density over a standard x86 server, the existing ClearSpeed product offers more like 27x, which is significant enough to warrant the effort to create a hybrid design that creates a far higher maximum performance.

Analysis

Before the advent of computing accelerators and the limits of power dissipation, maximizing the performance of a cluster simply meant buying the fastest x86 processors. Coprocessor accelerators have the ability to greatly increase performance per watt and performance per unit volume, and one can tune the ratio of coprocessors to general-purpose ("host") processors to achieve much higher performance within the facilities budget. We

now analyze the optimum use of these technologies.

Mixed integer-linear programming

Assume that a cluster consists of h host nodes, c ClearSpeed accelerators, and n future 64-bit NVIDIA accelerators (board set or 1U configuration). This creates a linear programming problem in the three dimensions h, c, and n. In addition to the linear constraint on power and weight and volume, there is an integer restriction as well: Each host node should have an integer number of accelerators as a replicated unit in the cluster, or else programming becomes exceedingly complicated. The goal of the mixed integer-linear programming problem is to maximize performance subject to the constraints.

When one component uniformly bounds the other, we can reject the inferior solution completely, thus simplifying the problem. Fortunately, this occurs in the case of ClearSpeed versus NVIDIA. The previous tables show far higher 64-bit performance per watt and per liter (and almost certainly, per kg). For example, a 1U x86 server with eight ClearSpeed boards has a peak speed of 730 GFLOPS. In contrast, a future 64-bit Tesla board set in a 1U x86 host would be between 128 GFLOPS and 172 GFLOPS, depending on how much the base 32-bit performance improves in the next few months. In addition, the NVIDIA solution has the lowest amount of memory per unit volume, thus reducing the capability of the system over either x86 or ClearSpeed hardware occupying the same space.

Figure 2 shows just the power constraint combined with the integer ratio constraint and the limit on PCIe slots per node. The solution space is the eight black dots, and the maximum performance is obviously the one corresponding to the 8:1 ratio.

The power restriction imposes a half-plane,

$$(h \times Ph) + (c \times Pc) < \text{Total power budget}$$

where Ph and Pc are the power consumed by each host node and each ClearSpeed accelerator, respectively.

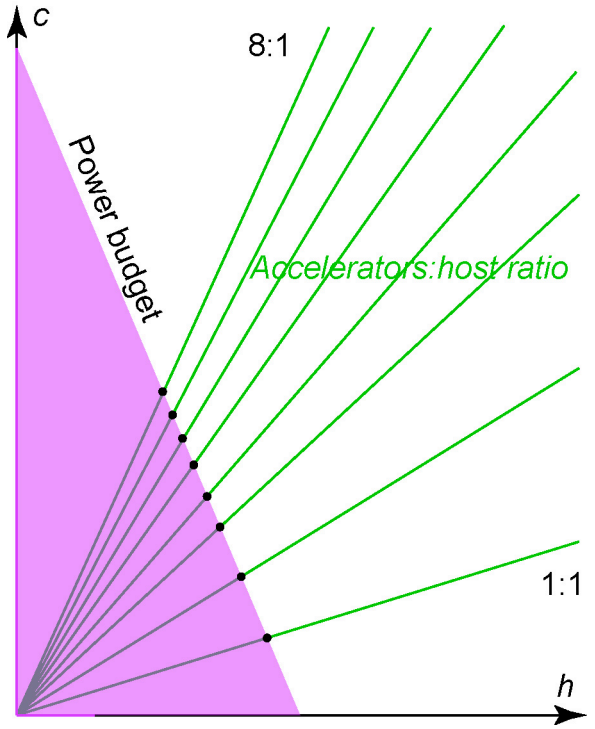

Figure 2. Integer-linear programming

If we accept $c = 8h$ in a maximum-performance design. We can now complete the example of the earlier section where the total power budget is 2 MW, of which 1.2 MW is available for the computing nodes themselves. If we use 700 watts for Ph and 35 watts for Pc, then

$$(h \times 700) + (c \times 35) < 1200000 \text{ W}$$

and $c = 8h$, which solves to

$$(h \times 700) + (c \times 35) \text{ watts}$$

$$= (700\,h + 280\,h) \text{ watts} < 1.2 \text{ megawatts}$$

$$\Rightarrow 980\,h < 1200000$$

which solves to $h < 1224$ nodes. In other words, 1224 nodes would completely use up all allocated electric energy. This corresponds to 34 racks of 36 servers each.

With 85.12 GFLOPS per x86 host and 80.64 GFLOPS per ClearSpeed board, each 1U server has a peak 64-bit speed of 0.73024 TFLOPS. The peak 64-bit multiply-add speed of such a cluster would be

$$1224 \times (85.12 + 80.64 \times 8) = 893.813 \text{ TFLOPS.}$$

Thus, one can build a nearly one-petaflops cluster within the 2 MW power budget coming into the building. The compute servers would occupy less than 400 square feet of floor space, which lessens latency and permits the use of very closely coupled communication technology. The total memory of such a system would be 50 TB, which is in line with the memory size balance of historical compute-intensive systems.

Any variation from this design to have more x86 and fewer ClearSpeed boards, or to use future NVIDIA board sets capable of 64-bit precision, will reduce the performance or exceed a constraint. Thus, we have established an example of maximum performance cluster design using mid-2007 hardware and facilities constraints. It is worth noting that each rack has a peak speed of 26.3 TFLOPS, which is substantially higher, say, than the 5 TFLOPS of a single Blue Gene/L cabinet from IBM. It is thus completely possible to have record compute density without leaving the world of mainstream Linux and x86 servers.

Summary

We have considered the design of clusters for maximum performance with three degrees of freedom: the number of x86 servers, the number of ClearSpeed accelerators, and the number of future 64-bit NVIDIA Tesla boards where we have used best available estimates for the performance of the latter.

Our conclusion is that future 64-bit Tesla boards, even in the most optimistic performance range, will have far less computing power per liter and per watt than current ClearSpeed boards. Since both depend on x86 hosts and since a host can only support about eight PCIe x8 slots, we conclude that one achieves the optimum performance density by using all eight PCIe slots for ClearSpeed cards within the same 1U enclosure that houses the x86 server. This configuration leads to very high computation density of about 26.3 TFLOPS per rack and 447 TFLOPS per facilities megawatt.