Low Power Amdahl Blades for Data Intensive Computing

Alex Szalay[1,3], Gordon Bell[2], Andreas Terzis[3], Alainna White[1] and Jan Vandenberg[1]

1. Dept. Physics and Astronomy, The Johns Hopkins University
2. Microsoft Research
3. Dept. of Computer Science, The Johns Hopkins University

2009-01-08

Abstract: The emergence of Solid State Disk (SSD) technology, poses the challenge of building a credible equivalent to a GrayWulf system with a similar IO performance, but with considerably lower power consumption. Alternatively an SSD-based system can offer significantly higher performance, while maintaining the same cost and power consumption. The results from our preliminary investigation are very encouraging (e.g., 12-fold decrease in total read time at the same cost or 7-fold decrease in power consumption with the same sequential IO), warranting a rapid evaluation of these ideas.

Introduction

The success of the GrayWulf (GW) architecture shows that an Amdahl number close to unity is a relevant metric to quantify what is an appropriate architectures for Data Intensive Scalable Computing. We have built a prototype of the GrayWulf system using commodity servers and a large number of relatively inexpensive SATA disks. However, looking a few years into the future, it is interesting to explore how to scale the different dimensions for high performance multi-petabyte storage systems. These dimensions include the Amdahl number, aggregate sequential IO speed, total time to read all the data, and total storage capacity. The total time to read through all the data is becoming an increasingly relevant issue for many data mining applications. While the capacity of normal hard disks is increasing rapidly, their access speed is only growing with the square root of the capacity. Thereby the time to read the whole disk is constantly increasing. One way to overcome this is to use more spindles, by installing smaller disks. However, in this case the power consumption grows rapidly, and one needs more controller channels. It is interesting to consider what would be the best way to build a modular system using SSDs, and what would be the different characteristics of such a clustered solution for data intensive applications.

Amdahl's Laws for a Balanced System

1. *Balanced system*: a system needs a bit of I/O per sec per instruction per sec;
2. *Memory*—the Mbyte/MIPS ratio (α) in a balanced system is 1.
3. *Input/output*—programs do one I/O per 50,000 instructions.

GrayWulf: the Reference Implementation

If we focus entirely on the sequential IO part, we needed about 24-30 SATA drives to reach a raw read performance of 1.5GB/s. The Tier-1 GW servers have 8x2.66GHz cores, resulting in an Amdahl number of 1.5*8/(8x2.66) = 0.56. The memory ratio for the GW is 0.75 MB/MIPS. Finally, the CPUS would require 426K IOPS, while our disks can only deliver about 6K IOPs, a shortage by a factor of 70.

The price for such a building block was about $12K, and has a total storage capacity of about 22.5TB. Its power consumption is approximately 1150W. Using $0.06 as the price of a kWh, the total energy cost over a three-year period is $1803. However, the current JHU rate is $0.15/kWh, and this raises the cost of electricity to $4,533 over three years.

The total time to read the all the disks at this speed is 15,000 sec. The GrayWulf consists of approximately 50 such servers, and this parallelism linearly increases the aggregate bandwidth, the total amount of storage and the power consumption. At the same time, the time to read all the disks remains constant.

Solid State Disks

Current Solid State Disks (SSDs) offer sequential IO speeds of approximately 90-250MB/s per disk (OCZ/Samsung/Intel). They are available today at a street price of around $300 for a 120 GB model, and $700-$900 for 256 GB. For the sake of an empirical comparison let us consider a nominal speed of 200MB/s. The newest SSDs (Intel X25-E, OCZ Vertex) are capable of 250MB/sec and about 35K IOPs[1,2]. These two disks appear to be the most cost effective today. While FusionIO has a higher sequential throughput [3] of 700 MB/s, at similar IOPs, its price is an order of magnitude higher then the previous two. Thus, considering a 256GB SSD with 200MB/s at a price of $300 is a reasonable projection for summer 2009.

Many SSDs today on the market have a power consumption similar to hard drives[4], but the OCZ/Samsung and the Intel drives represent an exception. They require about 0.2W of power idle, 2W at full speed.

There are two ways of using SSDs in a data-intensive computational and storage system. One is to use high-end servers and place multiple SSDs into the server, the same way we are building the GrayWulf nodes (scale-up). While this seems like the easiest way to go, our experiments show that the current high-end disk controllers saturate at around 740MB/sec, i.e. each 3 disks will require a separate expensive controller. Soon the servers will also run out of PCIe slots, and a matching networking bandwidth.

Scale Out: Low Power Motherboards

On the other hand, on large SQL Server systems there is already a trend to split the data into as many partitions as there are CPUs available. One should consider scale-out: use a separate CPU and host for each disk, build the ultimate brick that Jim Gray was always advocating. If we pair an SSD up with one of the recent low-power Intel CPUs (e.g. Atom N270/330), clocked at 1.6GHz we get an Amdahl number of 200x8 / 1,600 = 1.0. Various motherboards (ASUS EEE Box, Intel D945GCLF2, etc) are available using these CPUs.

Depending on which PCI chipset is used, the power consumption is between 20W-32W. Deploying such a motherboard equipped with 2GB of memory with one or two SSDs, one can build a “blade” with an Amdahl number close to 1, and a low power consumption of about 20W-40W. For the 1.6GHz Atom we would require 1.6E9/50,000 = 32K IOPs to have balanced system. Projecting a few months ahead into the future, one can see 256-512GB SSDs at the OCZ/Intel performance level emerging close to $300. The per disk access speed for now is probably not going to grow much, since the limiting factor will soon be the 3Gbits/s SATA/SAS bandwidth. Their power consumption is negligible compared to the motherboard, assuming one or two drives per blade. Finaly, the total time to read a 256 GB disk is 1,280s, a factor of 12 improvement over the GW.

A basic Amdahl Blade could consist of an ASUS eee box with an SSD, or an Intel D945GCLF2 motherboard with a dual core Atom N330 CPU and two SSDs. The latter configuration offers a higher packing density at the same level of relative power. Even an ASUS eee box with an SSD would beat the GrayWulf Amdahl number by a factor of 2. Our early experiments show that the ASUS+SSD power consumption is 16W average, 20W peak. A cluster of eight ASUS nodes will match the seqIO of the GW, while consuming less than 100W, compared to 1,150W for the GW. The IOPS would be a factor of 50 faster. The time to read through the disk would be 1,280s, a factor of 12 better than the GW. The only thing falling behind is the total storage capacity, with only 2TB vs. 22.5, a factor of 11.

Projecting a few months into the future, one can see 256GB SSDs emerging close to $300 bringing the total cost of an Eee blade to ~$600. So with $12K one could purchase 20 nodes, providing an aggregate sequential IO of 4GBytes/s, and a total storage of 4 TB, only a factor of 4 below the GW. If we consider an Intel D945GCLF2 motherboard, that has dual core Atom CPUs and can support two drives, one can build a more compact system, at roughly the same level of power consumption, but lower motherboard cost.

The estimated price at mid-summer 2009 for an ASUS-based system is about $600/node, $300 for the system, $300 for the SSD, projected for the summer of 2009. So from $12K we could buy 20 ASUS nodes with 1 SSD each, giving us a sequential IO of 4GBytes/s, and a total storage of 5TB, only a factor of 4 below the GW.

If we consider an Intel D945GCLF2 motherboard, that has dual core Atoms and can support two drives, one can build a more compact system, at roughly the same level of power consumption. We estimate the motherboard, memory and case to be about $200, and $600 for two SSDs, bringing the basic hardware to $800 for a dual node.

	CPU	*seq IO*	*randIO*	*disk*	*power*	*price*	*elect*	*cost*	*Amdahl*		*read*
	[GHz]	*[GB/s]*	*[IOPS/s]*	*[TB]*	*[W]*	*[$]*	*[$/3yr]*	*[$]*	*seq*	*IOP*	*[ksec]*
GrayWulf	21.28	1.5	6000	22.5	1150	12,000	4,533	16,533	0.56	0.014	15.00
ASUS	1.6	0.2	35000	0.256	20	600	79	679	1.00	1.094	1.28
Intel 330	3.2	0.4	70000	0.512	32	800	126	926	1.00	1.094	1.28

Table 1. Summary of the basic properties of the different node types, including the total cost of operating the system for 3 years, assuming $0.15/kWh.

Scaling the Low Power Amdahl Blade Systems

We consider a single GrayWulf node as our fiducial system, and scale the two low-power Amdahl Blades to match the GW performance along the different dimensions, cost (including the cost of electricity), total disk space, sequential IO and total power consumption.

From the data presented in Table 2 it is clear that even today, when SSD prices are still quite high, clusters of Amdahl blades are a very viable route to high performance IO. The first sub-table is particularly impressive, showing that at a fixed budget, we can improve the IO performance up to factors of 4.5 over the GrayWulf. The low power nodes only fall short in the amount of total disk space, but not by much.

It is interesting, that the power scaled to constant price is only a factor of 2-3 better than the GW. However, most of the power in the current motherboards is burnt up on the 945 chipset (PCI) and video subsystem. If one could design a special blade using a low power video (or have the video turned off). We can easily put 16 nodes into a 4-unit rack

COST	*nodes*	*seq IO*	*randIO*	*disk*	*cost*	*power*
		[GB/s]	*[kIOP/s]*	*[TB]*	*[$]*	*[W]*
GrayWulf	1.00	1.5	6.0	22.5	16,533	1150
ASUS	24.36	4.9	852.4	6.2	16,533	487
Intel 330	17.85	7.1	1249.6	9.1	16,533	571

DISK	*nodes*	*seq IO*	*randIO*	*disk*	*cost*	*power*
		[GB/s]	*[kIOP/s]*	*[TB]*	*[$]*	*[W]*
GrayWulf	1.00	1.5	6.0	22.5	16,533	1150
ASUS	87.89	17.6	3076.2	22.5	59,664	1758
Intel 330	43.95	17.6	3076.2	22.5	40,700	1406

SEQ IO	*nodes*	*seq IO*	*randIO*	*disk*	*cost*	*power*
		[GB/s]	*[kIOP/s]*	*[TB]*	*[$]*	*[W]*
GrayWulf	1.00	1.5	6.0	22.5	16,533	1150
ASUS	7.50	1.5	262.5	1.9	5,091	150
Intel 330	3.75	1.5	262.5	1.9	3,473	120

POWER	*nodes*	*seq IO*	*randIO*	*disk*	*cost*	*power*
		[GB/s]	*[kIOP/s]*	*[TB]*	*[$]*	*[W]*
GrayWulf	1.00	1.5	6.0	22.5	16,533	1150
ASUS	57.50	11.5	2012.5	14.7	39,033	1150
Intel 330	35.94	14.4	2515.6	18.4	33,283	1150

Table 2. The performance characteristics of a small cluster, equivalent to a single GW node, purchased mid 2009 from $12K. The different tables represent scaling to the same cost, total disk space, sequential IO and power consumption.

space, having 160 nodes in a single rack, with an aggregate IO speed of more than 30GBytes/sec, with 80TB of storage, consuming under 5kW.

Scaling to Thousands of Nodes

Considering their compact size and low heat dissipation, one can imagine building clusters of thousands of low-power Amdahl blades. In turn, this high density will create challenges related to interconnecting these blades using existing communication technologies (i.e., Ethernet, complex wiring if we have 10,000 nodes). On the other hand, current and upcoming high-speed wireless communications offer an intriguing alternative to wired networks. Specifically, current wireless USB radios (and their WLP IP-based variants) offer point-to-point speeds of up to 480 Mbps over small distances (~3-10 meters) [5]. Further into the future, 60 GHz-based radios promise to offer Gbps of wireless bandwidth [6].

While increasing aggressively, the capacity of wireless networking technologies will continue to trail that of wired alternatives. Furthermore, wireless bandwidth is inherently shared among all the nodes within the same broadcast domain. These characteristics mean that if large Amdahl blade cluster are to rely on wireless connectivity, the amount of data exchanged among cluster members need to be reduced. Interestingly, insights from wireless sensor networks (WSNs) and federated databases, which face a similar imbalance between processing capacity and network bandwidth, can be brought to bear. In both contexts, computation is moved closer to the raw data (i.e. individual sensing *motes* in the case of WSNs or individual DBs in the case of federated systems). Computation then distills raw data to more compact results than can be exchanged over the network. Fortunately, for scientific applications computation can either discard or aggregate enough data that at most 10% of the disk IO bandwidth is necessary. In concrete numbers, if one blade supports 200 MBps of disk bandwidth, the network needs to support 0.1*200*8 = 160 Mbps or network bandwidth, well within the capabilities of existing wireless technologies.

References

[1] http://download.intel.com/design/flash/nand/extreme/extreme-sata-ssd-datasheet.pdf
[2] http://www.ocztechnology.com/products/flash_drives/ocz_vertex_series_sata_ii_2_5-ssd
[3] http://www.fusionio.com/PDFs/Fusion%20Specsheet.pdf
[4] http://www.tomshardware.com/reviews/ssd-hard-drive,1968.html
[5] http://www.intel.com/technology/comms/uwb/download/Ultra-Wideband.pdf
[6] http://www.electronista.com/articles/08/02/22/nicta.gifi.chipset/