Low-Power Amdahl-Balanced Blades for Data Intensive Computing

Alexander S. Szalay[†] Gordon Bell[*] H. Howie Huang[‡] Andreas Terzis[†]
Alainna White[†]
Johns Hopkins University[†], George Washington University[‡], Microsoft Research[*]

ABSTRACT

Enterprise and scientific data sets double every year, forcing similar growths in storage size and power consumption. As a consequence, current system architectures used to build data warehouses are about to hit a *power consumption wall*. In this paper we propose a novel alternative architecture comprising large number of so-called *Amdahl blades* that combine energy-efficient CPUs with solid state disks to increase sequential read I/O throughput by an order of magnitude while keeping power consumption constant. We also show that while keeping the total cost of ownership constant, Amdahl blades offer five times the throughput of a state-of-the-art computing cluster for data-intensive applications. Finally, using the scaling laws originally postulated by Amdahl, we show that systems for data-intensive computing must maintain a balance between low power consumption and per-server throughput to optimize performance per Watt.

1. INTRODUCTION

Data sets generated by scientific instruments and business transactions continue to double per year, creating a dire need for a scalable data-intensive computing solution [3]. At the same time, the energy consumption of existing data warehouses increases linearly with their size, leading to prohibitive costs for building and operating ever growing data processing facilities [7]. The same observation motivated the JouleSort benchmark for evaluating the energy efficiency of computing platforms used for data-intensive applications [13]. The main challenge lies in the fact that existing systems used for data-intensive applications are *unbalanced*, whereby disk throughput cannot match CPU processing speeds and application requirements.

We propose to resolve this performance and energy efficiency conundrum by leveraging two recent technology innovations: Solid State Disks (SSDs) that combine high I/O rates with low power consumption and energy-efficient CPUs (e.g., Intel's Atom family of processors) originally developed for use in mobile computers. We show that it is possible to use these components to build balanced so-called *Amdahl blades* offering very high performance per Watt. Specifically, our experimental results show that Amdahl blade prototypes built using COTS components can offer five times the throughput of a current state-of-the-art data intensive computing cluster, while keeping the total cost of ownership constant. Alternatively, it is possible to keep the power consumption constant while increasing the sequential read I/O throughput by more than ten times.

2. BACKGROUND

Data-Intensive Computing. Scientific data sets are approaching petabytes today; enterprise data warehouses routinely store and process even more data. Most analyses performed over these datasets (e.g., data mining, regressions, aggregates and statistics) need to look at a large fraction of the stored data. Thereby, sequential read throughput is becoming the most relevant metric to measure the performance of data-intensive systems. Given that the relevant data sets do not fit in main memory, they have to be stored and retrieved from disks. For this reason, understanding the scaling behavior of hard disks is critical for predicting the performance of existing data-intensive systems as data sets continue to grow.

Over the last decade the rotation speed of large disks used in disk arrays has only changed by a factor of three, from 5,400 RPM to 15,000 RPM, while disk sizes have increased by a factor of 1,000. Likewise, seek times have improved only modestly over the same time period because they are limited by mechanical strains on the disk's heads. As a result, random access times have only improved slightly. Moreover, the sequential I/O rate continues to grow with the square root of disk capacity since it depends on the disk platter density [15].

As a concrete example of the trends described above, the sequential I/O throughput of commodity SATA drives is 60-80 MB/sec today, compared to 20 MB/sec ten years ago. However, considering the vast increase in disk capacity this modest increase in throughput has effectively turned the hard disk to a serial device: reading a terabyte disk at this rate requires 4.5 hours. Therefore, the only way to increase aggregate I/O throughput is to use more smaller disks and read from them in parallel. In fact, modern data warehouse systems, such as the GrayWulf cluster described next, aggressively use this approach to improve application performance.

GrayWulf. The GrayWulf system [16] represents a state-of-the-art architecture for data-intensive applications, having won the Storage Challenge at SuperComputing 2008. Focusing primarily on sequential I/O performance, each GrayWulf server consists of 30 locally attached 750 GB SATA drives, connected to two Dell PERC/6 controllers in a Dell 2950 server with 24 GB of memory and two four-core Intel Xeon processors clocked at 2.66 GHz. The raw read performance of this system is 1.5 GB/s, translating to 15,000 seconds (4.2 hours) to read all the disks. Such a building block costs approximately $12,000 and offers a total storage capacity of 22.5TB. Its power consumption is 1,150 W.

The GrayWulf consists of 50 such servers, and this parallelism linearly increases the aggregate bandwidth to 75 GB/sec, the total amount of storage to more than 1.1 PB and the power consumption to 56 kW. However, the time to read all the disks remains 4.2 hours, independent of the number of servers.

Doubling the storage capacity of the GrayWulf cluster, while maintaining its per-node current throughput, would require using twice as many servers, thereby doubling its power consumption. Alternatively, one could divide the same amount of data over twice as many disks (and servers) to double the system's throughput, at the cost of doubling its power consumption.

At this rate, the cost of building and operating these ever expanding facilities is becoming a major roadblock not only for universities but even for large corporations [7]. Thus tackling the next generation of data-intensive computations in a power-efficient fashion requires a radical departure from existing approaches.

3. AMDAHL-BALANCED BLADES

The previous discussion illustrates that a system's throughput is limited by its slowest component. Thereby for a given per-disk throughput, performance increases linearly with the total number of disks until the aggregate disk throughput saturates the CPUs' capacity for a given application workload. In practical terms, increasing the total number of disks requires increasing the number of servers, as the aggregate throughput of the locally-attached disk enclosure is configured to saturate the server's I/O bandwidth. At the same time, power consumption increases linearly with the number of servers. Having CPUs that can process data faster than the I/O subsystem can deliver is counterproductive: it does not increase the systems' throughput, while it increases its power consumption, an observation shared by Lim et al. for Internet computing workloads [10].

Gene Amdahl codified these relations in three laws that describe the characteristics of well-balanced computer systems [1]. Specifically, these laws state that a balanced computer system: **(1)** needs a bit of sequential I/O per sec per instruction per sec – *the Amdahl number*; **(2)** has memory with a Mbyte/MIPS ratio close to 1 – *the Amdahl memory ratio*; **(3)** performs one I/O operation per 50,000 instructions – *the Amdahl IOPS ratio*. The GrayWulf system has an Amdahl number of 0.56 and a memory ratio of 1.12. The third law requires 426 KIOPS to match the CPU speed, while hard disks can only deliver 6 KIOPS, a ratio of 0.014.

One can define the Amdahl number of computational problems as well: divide the size of the data set in bits with the cycles required to process it. Supercomputer simulations have Amdahl numbers of 10^{-5}, pipeline processing of observational astronomy data requires 10^{-2}, and user analyses of derived catalogs and database queries are close to unity. Thus, aiming for systems with high Amdahl numbers at a given performance level that match the Amdahl numbers of the applications is likely to result in balanced and thus energy-efficient systems.

Solid State Disks. Rather than increasing the number of disks, one should increase the per-disk throughput, thereby decreasing the number of servers, while keeping per-disk power consumption low. In fact, Solid State Disks (SSDs), that use similar flash memory as the one used in memory cards, provide both desired features.

Current SSDs offer sequential read I/O throughput of 90-250 MB/s and 10-30 KIOPS [9, 11]. The total time to read a 250 GB disk at these rates is 1,000 seconds, a factor of 15 improvement over the GrayWulf. These drives require 0.2W while idle and 2W at full speed [12]. They are available today at retail prices of $320 for a 120 GB model, and $600-$800 for 250 GB.

Projecting a few months into the future, the per disk sequential access speed is probably not going to grow considerably, since the current limiting factor is the 3 Gbit/s SATA bandwidth. Further ahead, the emergence of 6 Gbit/s SATA controllers on inexpensive motherboards and SSDs will provide a way to higher sequential speeds at an affordable price point. The only other way to exceed this limitation is to put the flash mem-

Table 1: Low-power systems used in our comparison.

System	Model	CPU	Chipset
ASUS	EeeBox	N270	945GSE
Intel	D945GCLF2	N330	945GC
Zotac	Ion	N330	ION
AxiomTek	Pico 820	Z530	US15W
Alix	3C2	LX800	AMD

ory directly onto the motherboard, eliminating the disk controller. The market will probably force motherboard and disk manufacturers to stay with the standard SATA interfaces for a while to ensure large production quantities and economies of scale. We believe that boutique solutions with a direct access to flash, such as the FusionIO products [6], are unlikely to become a commodity.

Scale-up: SSDs on High-end Servers. One way to deploy SSDs in data-intensive computations is through an approach we term *scale-up:* use high-end servers and connect multiple SSDs to each server, the same way we have built the GrayWulf nodes. While this appears to be the most intuitive approach, our experiments show that current high-end disk controllers saturate at 740 MB/sec. In turn, this limit means that each set of three SSDs require a separate controller. Soon servers will run out of PCI slots as well as PCI throughput.

Scale-down: Low Power Systems. Instead, we take the current trend of splitting data into multiple partitions across multiple servers [5], to its logical extreme: use a separate CPU and host for each disk, building the cyberbrick originally advocated by Jim Gray [2].

In fact, if we pair an SSD with one of the recent energy-efficient CPUs used in laptops and netbooks (e.g., Intel's 1.6GHz Atom N270 [8]), we arrive at an Amdahl number close to one. Moreover, the IOPS Amdahl ratio is very close to ideal: a 1.6 GHz CPU would be perfectly balanced with 32,000 IOPS, close to what current SSDs can offer. Given its balanced performance across all the dimensions mentioned in Amdahl's laws, we term such a server an *Amdahl blade*. Adding a dual-core CPU and a second SSD to such a blade increases packing density at a modest increase in power since the SSDs consume negligible power compared to the motherboard.

4. EVALUATION

We built prototypes of such Amdahl blades using COTS components to evaluate their potential in data-intensive applications.

Table 1 compares the characteristics of the systems used in our tests. All Amdahl blades use variants of the Intel Atom processor clocked at 1.6 GHz. The N330 CPU has two cores while the rest have a single core. We compare them to the GrayWulf system [16] and the Alix 3C2 node that uses the LX800 500 MHz Geode CPU from AMD and a Compact Flash (CF) card for storage. We include the Alix node in our comparison because it is used by the FAWN project that recently proposed an alternative power-efficient cluster architecture for data-intensive computing [17]. Rivoire et al. have previously investigated the energy-saving benefits of a configuration similar to the Amdahl blade in the context of an external sorting benchmark [14]. Our results reflect recent advances in SSD performance and include total cost of ownership as one of the comparison metrics.

We experimentally measure the blades' performance by installing Windows 7 Release Candidate and running the SQLIO utility that simulates realistic sequential and random disk access patterns [4]. We vary block size from 8 KB to 1 MB at 4x increments. Furthermore, we run each test using 1, 2, and 32 threads. Each test runs for sixty seconds using an 8 GB dataset. We use previously reported measurements for the Alix system assuming an 8 GB CF card [17], while the GrayWulf was previously evaluated using a similar methodology [16].

We measure power consumption under peak load, using both a Kill-A-Watt power meter and directly at the DC input of the motherboards, whenever possible.

Throughput and Power Consumption. The CPU column in Table 2 corresponds to the individual CPU speed multiplied by the number of cores. While this metric overlooks important performance aspects, such as differences in CPU micro-architectures and available level of parallelism, we use it as a first approximation of processing throughput used to calculate the relative Amdahl numbers. We use one SSD per core and therefore the Intel and Zotac motherboards that utilize the same two-core Intel Atom N330 CPU have two drives. All SSD tests use identical OCZ 120 GB Vertex drives [11].

The Zotac and Intel boards offer the best sequential read performance, 250 MB/s per SSD or an aggregate of 500 MB/s, using two threads. This value was obtained for block sizes of 256 KB, due to the Atom's 512 KB L1 cache. The aggregate sequential read rate decreases to 450 MB/s with 32 threads on the dual-core motherboards. On the other hand, the maximum sequential I/O for single-core motherboards is only 124 MB/s. Furthermore, the maximum per disk write performance levels off at 180 MB/s for random I/O and 195 MB/s for sequential I/O. Finally, the dual-core boards deliver 10.4 KIOPS compared to 4.4 KIOPS for the single-core boards under a workload of random read patterns.

We calculate the total cost of ownership by estimating the cost of purchasing and operating each system over a period of three years. We calculate the acquisition cost using current (06/09) retail prices for motherboards and the actual prices used to purchase the GW system in July

Table 2: Performance, power, and cost characteristics of various data-intensive architectures.

	CPU	Mem	SeqIO	RandIO	Disk	Power	Cost	Relative	Amdahl numbers		
	[GHz]	[GB]	[GB/s]	[kIOPS]	[TB]	[W]	[$]	Power	Seq	Mem	Rand
GrayWulf	21.3	24	1.500	6.0	22.5	1,150	19,253	1.000	0.56	1.13	0.014
ASUS	1.6	2	0.124	4.6	0.25	19	820	0.017	0.62	1.25	0.144
Intel	3.2	2	0.500	10.4	0.50	28	1,177	0.024	1.25	0.63	0.156
Zotac	3.2	4	0.500	10.4	0.50	30	1,189	0.026	1.25	1.25	0.163
AxiomTek	1.6	2	0.120	4.0	0.25	15	995	0.013	0.60	1.25	0.125
Alix 3C2	0.5	0.5	0.025	N/A	0.008	4	225	0.003	0.40	1.00	

2008 (essentially the same today). We note that for the SSD-based systems the cost and disk size columns in Table 2 represent projections for a 250 GB drive with the same performance and a projected cost of $400 at the end of 2009, in line with historic SSD price trends.

Power consumption varies between 15W-30W depending on the chipset used (945GSE, USW15, ION) and generally agrees with the values reported in the motherboards' specifications. The current university rate for electric power at JHU is $0.15/kWh. The total cost of power should include the cost for cold water and air conditioning, thus we multiply the electricity cost by 1.6 [7]. Table 2 presents these cumulative costs.

Lastly, we present the different Amdahl numbers and ratios for the various node types. It is clear that, compared to the GrayWulf and Alix, the Atom systems, especially with dual cores, are better balanced across all three dimensions.

Scaling Properties. Table 3 illustrates what happens when we scale the other systems to match the GrayWulf's sequential I/O, power consumption, and disk space. We present the number of nodes necessary to match the GW's performance in the selected dimension, while the remaining columns show the aggregate performance across all these nodes.

We note that a cluster of only three Intel or Zotac nodes will match the sequential read I/O of the GrayWulf and deliver five times faster IOPS, while consuming 90W, compared to 1150W for the GW. The only shortcoming of this alternative is that the total storage capacity is 15 times smaller. At the same time, the power for a single GrayWulf node can support 41 Intel and 38 Zotac nodes, respectively, and offer more than ten times higher sequential read I/O throughput.

Table 3 also shows that one needs to strike a balance between low power consumption and high performance. For example, while the sequential read I/O performance of the Alix system matches that of the GrayWulf at a constant price, it falls behind that of the Amdahl blades. Furthermore, one needs 60 Alix boards to match the sequential rate of a GW node which consume approximately three times more power than the equivalent Intel system (240 W vs. 84 W).

5. DISCUSSION

The nature of scientific computing is changing – it is becoming more and more data-centric while at the same time datasets continue to double every year, surpassing petabytes. As a result, the computer architectures currently used in scientific applications are becoming increasingly energy inefficient as they try to maintain sequential read I/O performance with growing dataset sizes. The scientific community therefore faces the following dilemma: find a low-power alternative to existing systems or stop growing computations on par with the size of the data. We thus argue that it is unavoidable to build *scaled-down and scaled-out* systems comprising large numbers of compute nodes each with a much lower relative power consumption at a given sequential read I/O throughput.

We use Amdahl's laws to guide the selection of the smallest CPU throughput necessary to run data-intensive workloads dominated by sequential reads. Furthermore, we propose a new class of so-called *Amdahl blades* that combine energy-efficient processors and solid state disks to offer significantly higher throughput and lower energy consumption. We find that today the dual-core Amdahl blades represent a sweet spot in the energy-performance curve, while alternatives using lower power CPUs (i.e., single-core Atom, Geode) and Compact Flash cards offer lower relative throughput. As technology trends evolve, we believe that Amdahl's laws can continue to guide the design of servers in the future.

The only advantage of existing systems is their higher total storage space. However, as SSD capacities are undergoing an unprecedented growth, this temporary advantage will rapidly disappear: as soon as we have a 750 GB SSD for $400, the storage built of low-power systems will have a lower total cost of ownership than regular hard drives. An intriguing alternative is using nodes in which one SATA port will be connected to an SSD while the other port(s) will be connected to low-

Table 3: Comparison of the systems scaled to various dimensions.

	CPU [GHz]	SeqIO [GB/s]	RandIO [kIOPS]	Disk [TB]	Power [W]	Cost [$]	Relative Power	Nodes
				Scaled to constant total cost				
GrayWulf	21.3	1.5	6	22.5	1150	*19250*	1.000	1
ASUS	37.6	2.9	108	5.9	446	*19250*	0.388	23.5
Intel	52.4	8.2	164	8.2	458	*19250*	0.398	16.4
Zotac	51.8	8.1	168	8.1	486	*19250*	0.422	16.2
AxiomTek	31.0	2.3	77	4.8	290	*19250*	0.252	19.4
Alix 3C2	42.7	2.1	N/A	0.7	342	*19250*	0.297	85.5
				Scaled to constant sequential read				
GrayWulf	21.3	*1.5*	6	22.5	1150	19250	1.000	1
ASUS	19.4	*1.5*	56	3.0	230	9920	0.200	12
Intel	9.6	*1.5*	30	1.5	84	3530	0.073	3
Zotac	9.6	*1.5*	31	1.5	90	3570	0.078	3
AxiomTek	20.0	*1.5*	50	3.1	188	12430	0.163	12.5
Alix 3C2	30.0	*1.5*	N/A	0.5	240	13510	0.209	60
				Scaled to constant power				
GrayWulf	21.3	1.5	6	22.5	*1150*	19250	1.000	1
ASUS	96.8	7.5	278	15.1	*1150*	49620	1.000	60.5
Intel	131.4	20.5	411	20.5	*1150*	48330	1.000	41.1
Zotac	122.7	19.2	399	19.2	*1150*	45590	1.000	38.3
AxiomTek	122.7	9.2	307	19.2	*1150*	76250	1.000	76.7
Alix 3C2	143.8	7.2	N/A	2.3	*1150*	64750	1.000	287.5
				Scaled to constant disk space				
GrayWulf	21.3	1.5	6	*22.5*	1150	19250	1.000	1
ASUS	144	11.3	414	*22.5*	1710	73790	1.500	90
Intel	144	22.5	450	*22.5*	1260	52950	1.100	45
Zotac	144	22.5	468	*22.5*	1350	53520	1.200	45
AxiomTek	144	10.8	360	*22.5*	1350	89515	1.200	90
Alix 3C2	1406	70.3	N/A	*22.5*	11250	633460	9.800	2812

power conventional disks.

While offering unprecedented performance, the proposed architecture also introduces novel challenges in terms of data partitioning, fault tolerance, and massive computational parallelism. Interestingly, some of the approaches proposed in the context of wireless sensor networks and federated databases, that advocate keeping computations close to the data, can be translated to this new environment.

Acknowledgments

The authors would like to acknowledge financial support from Microsoft External Research, the Gordon and Betty Moore Foundation, and the NSF.

6. REFERENCES

[1] G. Amdahl. Computer architecture and amdahl's law. *IEEE Solid State Circuits Society News*, 12(3):4–9, 2007.

[2] T. Barclay, W. Chong, and J. Gray. Terraserver bricks: A high availability cluster alternative. Technical Report MSR-TR-2004-107, Microsoft Research, 2004.

[3] G. Bell, T. Hey, and A. Szalay. Beyond the data deluge. *Science*, 323(5919):1297–1298, 2009.

[4] D. Cherry. Performance Tuning with SQLIO. Available from: `http://sqlserverpedia.com/wiki/SAN_Performance_Tuning_with_SQLIO`, 2008.

[5] P. Furtado. Algorithms for Efficient Processing of Complex Queries in Node Partitioned Data Warehouses. *Database Engineering and Applications Symposium, 7-9 July*, pages 117–122, 2004.

[6] Fusion-IO. ioDrive. Available from: `http://www.fusionio.com/`.

[7] J. Hamilton. Cooperative expendable micro-slice servers (cems). In *Proceedings of CIDR 09*, 2009.

[8] Intel. Intel Atom Processor. Available from: `http://www.intel.com/technology/atom/`, 2009.

[9] Intel Corporation. Intel x25-e sata solid state drive. Available from: `http://download.intel.com/design/flash/nand/extreme/extreme-sata-ssd-datasheet.pdf`.

[10] K. Lim, P. Ranganathan, J. Chang, C. Patel, T. Mudge, and S. Reinhardt. Understanding and designing new server architectures for emerging warehouse-computing environments. In *ISCA '08: Proceedings of the 35th International Symposium on Computer Architecture*, pages 315–326, Washington, DC, USA, 2008. IEEE Computer Society.

[11] OCZ Technology. OCZ Flash Media: OCZ Vertex Series SATA II 2.5 SSD. Available from: `http://www.ocztechnology.com/`.

[12] P. Schmid and A. Roos. Flash SSD Update: More Results, Answers. Available from: `http://www.tomshardware.com/reviews/ssd-hard-drive,1968.html`, 2008.

[13] S. Rivoire, M. Shah, P. Ranganathan, and C. Kozyrakis. Joulesort: a balanced energy-efficiency benchmark. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 365–376, New York, NY, USA, 2007. ACM.

[14] S. Rivoire, M. Shah, P. Ranganathan, C. Kozyrakis, and J. Meza. Models and Metrics to Enable Energy-Efficiency Optimizations. *IEEE Computer*, 40(12):39–48, Dec. 2007.

[15] S. Sankar, S. Gurumurthi, and M. R. Stan. Intra-disk parallelism: An idea whose time has come. *SIGARCH Comput. Archit. News*, 36(3):303–314, 2008.

[16] A. Szalay and G. Bell et al. GrayWulf, Scalable Clustered Architecture for Data Intensive Computing. In *Proc. of HICSS-42 Conference*, 2009.

[17] V. Vasudevan, J. Franklin, D. Andersen, A. Phanishayee, L. Tan, M. Kaminsky, and J. Moraru. FAWNdamentally Power Efficient Clusters. In *Proceedings of HotOS*, 2009.