Challenges in Building a Flat-Bandwidth Memory Hierarchy for a Large-Scale Computer with Proximity Communication

Robert Drost, Craig Forrest, Bruce Guenin, Ron Ho, Ashok V. Krishnamoorthy, Danny Cohen, John E. Cunningham, Bernard Tourancheau, Arthur Zingher, Alex Chow, Gary Lauterbach, and Ivan Sutherland*

Abstract

Memory systems for conventional large-scale computers provide only limited bytes/s of data bandwidth when compared to their flop/s of instruction execution rate. The resulting bottleneck limits the bytes/flop that a processor may access from the full memory footprint of a machine and can hinder overall performance. This paper discusses physical and functional views of memory hierarchies and examines existing ratios of bandwidth to execution rate versus memory capacity (or bytes/flop versus capacity) found in a number of large-scale computers. The paper then explores a set of technologies, Proximity Communication, low-power on-chip networks, dense optical communication, and Sea-of-Anything interconnect, that can flatten this bandwidth hierarchy to relieve the memory bottleneck in a large-scale computer that we call "Hero."

1. Introduction

High Performance Technical Computing (HPTC) is driven by rising computational demands from private industry and government sectors. The processing requirements for many HPTC applications outstrip single-chip and modest multi-chip (*e.g.*, 4- or 8-way) processors and rely on increasingly massive parallel computer system architectures. Such architectures require high bandwidth communication in order to effectively utilize their increasing number of processors and memory subsystems.

Currently, the top ten High Performance Computer (HPC) systems have system bisection bandwidths between 1 and 8TB/s [12]-[35]. A next-generation system with a 100-fold improvement in bandwidth would enable significant advances in large-scale computer performance and programmer productivity. We are investigating a communication platform for such a system, "Hero."[1] The Hero communication platform increases system-level interconnect density by 100-fold, providing a corresponding increase in system throughput over today's machines. Hero contains modules based on Proximity Communication along with power-efficient on-chip networks to connect a grid of chips with high bandwidth. Modules are interconnected by massively parallel optical links.

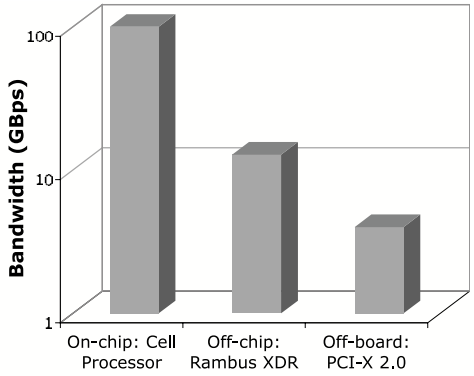

Figure 1. Example of bandwidths at various physical levels in 2005

This paper is organized as follows. In Section II we consider the memory bottleneck as found in circa 2005 systems. Then from physical and functional viewpoints, Sections III - VI explore three communication technologies: Proximity Communication, low-power on-chip signaling, and dense optical communication. These increase off-chip, on-chip, and off-module bandwidths respectively. In Section VII we present a distributed switching transport layer that transparently carries data through the three physical communication technologies.

2. The Memory Bottleneck

Present computer systems suffer from a performance-limiting memory bottleneck. The "bottle" arises from the bandwidth between computing elements and various quantities or levels of memory. For example, in most systems, datapaths see much higher bandwidth to register files and first-level caches than to the channels that bring data from distant main memory. Taking a reverse view, memory cells see much higher bandwidth to on-chip buses than to the channels that carry their data to distant datapaths.

*This work was supported in part by DARPA as part of its HPCS Phase II program, NBCH3039002.

[1] "Hero" is not an acronym, but rather was coined by Ivan Sutherland because such a system takes on heroic proportions.

Proceedings of the 13th Symposium on High Performance Interconnects (HOTI'05)
1550-4794/05 $20.00 © 2005 IEEE

This relationship between bandwidth and memory can be examined from physical and functional viewpoints. A physical classification of the memory hierarchy may contain three principal communication levels: on-chip, off-chip, and off-board. Figure 1 shows high-performance bandwidth examples: an on-chip Sony-IBM-Toshiba Cell processor inter-core network, an off-chip Rambus XDR memory interface, and an off-board PCI-X 2.0 interface [1]-[4]. The steep steps, or "cliffs," from on-chip to off-chip to off-board bandwidths have historically been a dominating force in computer architecture.[2]

The first physical communication level utilizes on-chip wires. Qualitatively, on-chip wires are often characterized as limiting chip performance. In fact, on-chip wires provide immense cross-section bandwidth across a VLSI chip [5] and enable multitudes of high performance computing and memory cells to communicate across a chip.

The second physical communication level is off-chip. Off-chip bandwidth is the product of the number of off-chip channels and the communication rate per channel. The number of off-chip channels in high-performance packages may increase at a rate of only about 9% per year over the next decade [6]. Given this limited increase, work has been done to increase data rates using, for example, high speed serial transceivers [7],[8]. However there remains a widening disparity between on-chip cross-section bandwidth and off-chip bandwidth.

The third physical communication level is board-to-board. Physical channels may use rigid perpendicular boards connecting through connectors to a backplane, flexible channels of twisted-pair or coaxial wires, or fiber optic channels in bundled or ribbon configurations. Historically, electrical connectors provide higher connection counts than optical; however, lower attenuation in optical channels gives them significant distance and bandwidth advantages over "long" wires. As optical technology costs come down, the definition of "long" has reduced from kilometers to meters. Current research has explored using optics at the board and even the chip level, although the cost-benefit ratio for optics is presently less compelling than that for conventional electrical technologies at those levels.

Next, let us consider a functional view of the memory hierarchy. Figure 2 shows an example hierarchy for a scalar architecture: registers, first, second, and third level caches, and memory on nearby and distant boards. Successive functional memory levels typically exhibit increasing capacity and latency, and decreasing bandwidth. The physical location and presence or absence of layers in a given system depends on its scale and memory reference patterns. From the viewpoint of an execution pipeline, each level of memory has some bandwidth, capacity, and latency where these properties are strongly interdependent.

[2]For instance, although increasing chip size incurs a disproportionate increase in cost due to wafer defects, chips have grown to include multiple caching levels and higher pin counts in order to compensate for these cliffs.

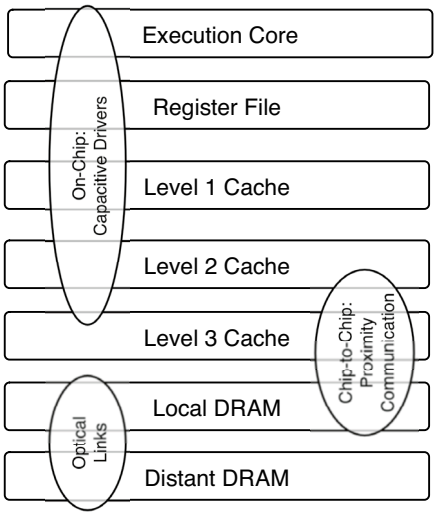

Figure 2. Functional and physical level relationship for a scalar architecture

Memory system design tradeoffs depend on the different bandwidth and caching needs of scalar, vector, and stream architectures. For instance, second and third level caches have been popular in scalar systems since the early 1990s due to the inability of first level caches to successfully hide memory latency and bandwidth limitations. In contrast, vector and stream architectures lead to less data re-use; hence such additional caching can hinder memory transfers. Caching is limited or absent in these architectures. Instead, local vector registers or stream buffers are used.

A general goal of the functional memory hierarchy, whether it uses scalar caches, vector registers, or stream buffers, is to reduce bandwidth requirements to physically distant memory. However, some applications, modeled by benchmarks such as Tabletoy or RandomAccess that measure performance in giga-updates-per-second (GUPS), provide little or no opportunity for data reuse [9],[10]. Therefore, these applications require high bandwidth to the full memory footprint of a machine. Furthermore, even when an application potentially contains spatial and temporal data locality, the burden of finding and exploiting these localities can reduce programmer productivity [11].

One measure of the balance between memory and processing in a HPC system is the ratio of bandwidth to execution rate, or $bytes/s$ to $flop/s$, or the timeless ratio of $bytes/flop$. An execution core sees differing bytes/flop ratios to the various levels of a memory hierarchy. For instance, a first-level cache typically serves a sole execution core with high bandwidth whereas many cores may share a third-level cache and, of course, all execution cores share the full memory footprint of a machine.

This $bytes/flop$ ratio will vary across memory capaci-

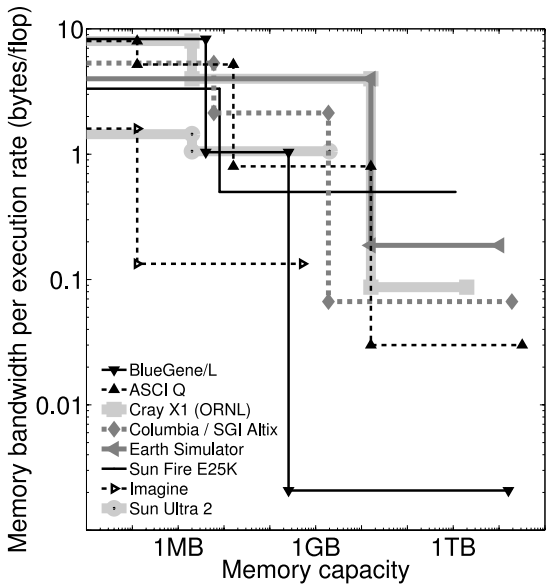

Figure 3. Bandwidth versus memory capacity

ties. Figure 3 plots the $bytes/flop$ versus memory size for a set of (mostly) large-scale computers [12]-[35].[3] The data set includes two vector machines (Earth Simulator and Cray X1), a number of superscalar machines (Sun Fire E25K, Sun Ultra 2, SGI Altix, ASCI Q, and BlueGene/L), and a streaming machine (Imagine). Each cliff corresponds to a point where the execution cores must communicate with a more distant level of the memory hierarchy in order to access additional capacity. In constructing each profile, we assume the system is installed with maximum capacity at each memory level that is supported or addressable by the system, except for installed systems for which the actual memory capacities in known.

For the full memory footprint of a machine all execution cores share that memory. Hence the right side of the graph shows bisection bandwidth divided by the peak flops for the machines–a number that is usually reported. At the left edge of the graph individual execution cores access their first level caches. Processor architects typically design this level of memory to have sufficient bandwidth to feed near-peak flop performance. As a result, the data sets have a narrower spread on the left rather than right side of the graph.

For intermediate memory capacities we ratio the total bandwidth to peak flops for the execution cores sharing memory. For example, if 16 cores each provide 4 $Gflop/s$ and share a 256MB third-level cache with 64 $Gbytes/s$ of bandwidth, then the $bytes/flop$ would equal $64/16 * 4$, or 1, for this 256MB memory capacity.

[3]Most of the featured systems are recent installations; naturally, their configurations and performance will change over time. Although our data reflects the systems in their reported states, upgrades and installations of additional nodes will only extend the tails in memory capacity without affecting the downward staircase profile of $bytes/flop$ at the various levels of the memory hierarchy.

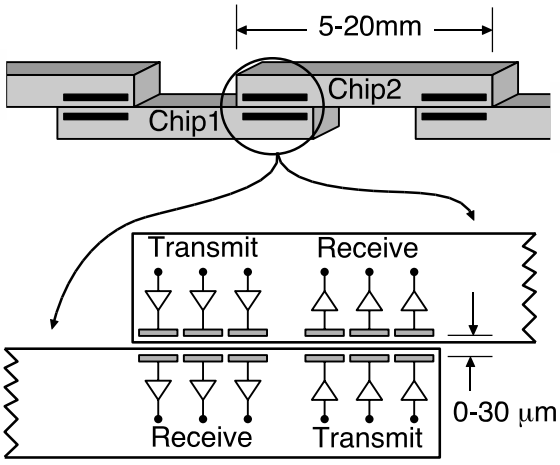

Figure 4. Proximity Comm. cross-section

Although this paper focuses on the challenges of increasing bandwidth, latency can also be critical to computer system performance. Figure 3 only addresses memory performance in terms of bandwidth, not latency; however, for most systems latency climbs correspondingly upward as we move to more distant memory levels. To accomodate latency, an application must still perform useful work while delayed memory references are in flight. For example, given a latency, L, and cycle rate, C, then a minimum of L/C memory references must be simultaneously in flight while code performs other useful work or else this latency will limit computing throughput. Although we shall not discuss the details in this paper, Proximity Communication and our highly integrated optical communication permit smaller physical structures to connect execution cores and memory. Because the speed of light provides lower latency for smaller structures, our system can also provide lower latencies.

The remaining sections of this paper explore technologies that have the potential to flatten the curves in Figure 3 by raising their right sides. This exploration reveals some of the possibilities and challenges for building memory systems with $bytes/flop$ to the most distant gigabytes of memory that can equal the $bytes/flop$ to a first level cache.

3. Proximity Communication chip-to-chip

Proximity Communication aims to solve the off-chip bandwidth bottleneck. We have presented some key technology features of Proximity Communication in prior publications [36, 37]. Therefore, this section first summarizes the key features and challenges of Proximity Communication and then presents a new analysis of its bandwidth sensitivities and limitations.

Figure 4 shows a cross-section view of chips using Proximity Communication, in which their face-to-face place-

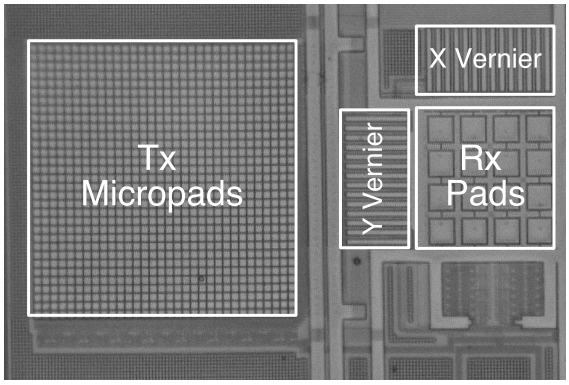

Figure 5. Electronic alignment chip photo

ment allows communication via capacitive coupling of top layer metal pads. Replacing off-chip high-speed wires with Proximity Communication reaps a number of benefits. First, the density of channels follows lithography improvements and thus scales up with Moore's Law. Second, the power per channel is less than conventional chip I/O because transmit and receiver circuits are small, the capacitive channel requires no equalization for inter-symbol interference, and the pads are covered by overglass, obviating high-capacitance electro-static discharge protection circuits. Third, Proximity Communication avoids permanent attachment of face-to-face chips and permits replacement of defective chips. This solves the known-good-die problems found in other multi-chip integration technologies such as multi-chip modules or system-in-package.

The key challenge is alignment. We have discussed how electronic alignment can correct for in-plane mechanical misalignments in X, Y, and Θ using transmit micro-pads [37]. Figure 5 shows a chip plot of transmit micro-pads that correct misalignment by steering data to the receiver pads. However, out-of-plane misalignment in Z, Ψ, and Φ cannot be corrected by a similar mechanism, and hence set a limit on the maximum tolerable misalignment.

A simplified equation for Proximity bandwidth is

$$BandwidthDensity = \frac{Gb/s}{channel} * \frac{channels}{area} * \frac{1}{overhead}$$

Optimized, high-power serial links can transmit and receive data at a bit period of about one fanout-of-four inverter delay ($FO4$) [7]. In contrast, lower-speed chip I/O uses a bit period comparable to chip clock periods of 12 to 20 $FO4$ or more. Balancing link speed and energy efficiency leads to bit periods of about 4 to 6 $FO4$ delays [38]. For a 90-nm technology with a $FO4$ delay of about 30ps, a bit period of 5 $FO4$ delays provides about 6 $Gb/s/channel$.

Figure 6 shows a cross-section view of pads on two chips, ChipA and ChipB, that communicate using capacitive coupling with a pitch, P, between adjacent pads and a gap separation, G, between pads on the two chips. For simplicity, transmit pads rather than micropads are

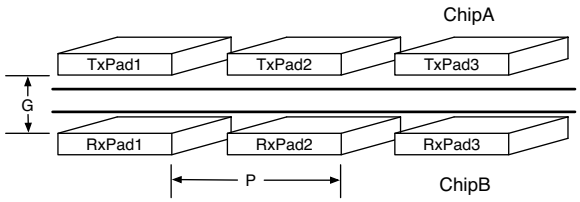

Figure 6. Gap and pitch of Proximity pads

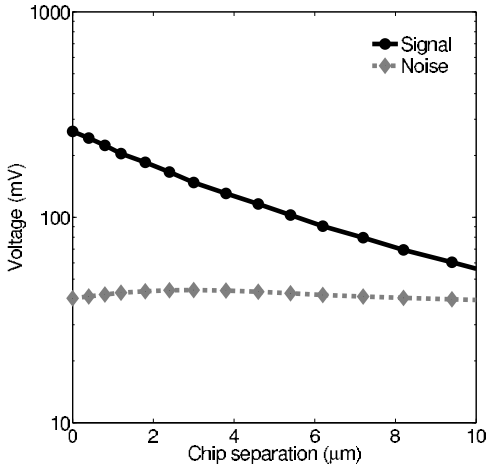

Figure 7. Simulated signal and noise coupling

shown. Vertically aligned pads couple capacitively to communicate whereas adjacent transmit and receive pads introduce cross-coupling noise. The density per area of channels is $1/(2P^2)$. As the gap G increases, the received signal decreases due to reduced signaling capacitance and noise increases due to greater cross-coupling capacitance, as shown in Figure 7 for 20 micron pitch differential pads pitch with worst case X and Y misalignment.

For any pitch P there exists some maximum gap G at which the receiving circuit can no longer sense the signal, yielding a misalignment constraint of $G \leq \alpha P$ and hence $\alpha^2/(2G^2)$ channels per unit area. α ranges from 0.1 to 1.0 depending on factors such as receiver sensitivity and dielectric permittivity, and a typical value for α is 0.2. Overhead consists of extra pads for clocking and to digitally compensate for X, Y, and Θ misalignment. Both overheads can be estimated at about 25%, or about 1.56x combined.

Merging these estimates, and using $\alpha = 0.2$ gives

$$BandwidthDensity = \frac{77}{G^2}\ \frac{Tb/s}{mm^2}$$

This equation shows that the bandwidth can be quite high, but it depends on the inverse square of the gap. For example, using 10% of a 150mm^2 chip for Proximity Communication with a gap misalignment of under 5 microns yields a bandwidth per square millimeter of about 3.1 Tb/s and a chip bandwidth of about 46 Tb/s. Such a chip would easily match or exceed on-chip bandwidths.

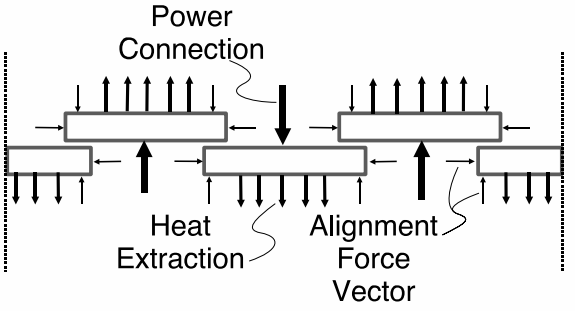

Figure 8. Possible alignment, thermal, and power mapping onto chips in an array

4. Packaging Challenges

Proximity Communication enables unprecedented communication bandwidth but requires packing large numbers of chips in a dense array, as in Figures 8. This compact array of chips poses significant chip packaging challenges:

- **Mechanical:** Maintain the required alignment between chips, despite fabrication and assembly tolerances, as well as thermal and mechanical perturbations encountered during operation.
- **Thermal:** Extract several kiloWatts (kW) of heat from a module while maintaining all devices within their allowed temperature range.
- **Power:** Provide kW of power in the form of kiloAmps of current while limiting ohmic losses and $\frac{dI}{dt}$ transients in the power distribution.

Low-power applications such as optics routinely use electronic components aligned within fine tolerances [39]. However, achieving these alignment tolerances in the presence of large-scale power and current requirements poses some unique challenges. Figure 8 illustrates a possible mapping of the alignment, thermal, and power functions onto an array of chips using Proximity Communication. Our packaging solution uses a precision structure to force each chip into X, Y, and Θ alignment with a global coordinate system centered on the chip array. While this would still allow random variations in the X, Y, and Θ position of each chip, it would prevent misalignment from accumulating from one end of the chip array to the other. This precision alignment structure should be engineered to maintain adequate dimensional stability, even as the chips in the array heat and cool with varying computation and communication workloads.

The allowed gap between facing chips is set by the Proximity Communication circuits. Maintaining this gap between chips requires an additional planarity constraint of the individual chips, as well as a compliant structure to force the two chip layers into contact or near contact. The success of the alignment structure requires that the forces pushing the chips into alignment dominate any misaligning forces resulting from the physical connections between the chips and the thermal and power distribution support structures.

Thermal challenges arise in arrays of over 100 chips using Proximity Communication, dissipating in excess of 10kW and with heat flux levels ranging from 10 to 50 W/cm^2, depending on the composition and placement of the chips. The optimal choice for managing these heat flow densities uses a water-cooled cold plate. With appropriate manifold design, each chip site can be cooled with water which has not been pre-heated by the other chips [40]. This not only provides efficient cooling, but also minimizes lateral thermal gradients from one end of the cold plate to another, hence reducing any thermal distortions of the cold plate which may impact chip alignment. The thermal interface material between each chip and the cold plate surface plays an important role in the thermal design. It must provide a stable thermal resistance between the mating chip and cold plate services [41], and should be applied with adequate thickness control in order to maintain the planarity of the chips in the array.

Power conversion circuitry and the power distribution hardware place demands that are, in principle, no different from those of other computer systems. A successful power distribution system will present a low impedance to voltage sources over many decades of frequency. This is accomplished by putting the voltage conversion circuitry as close as possible to the load; ensuring adequate metal in the current paths to control ohmic losses; positioning the Vdd and ground paths close to each other to minimize current loop inductances; and judiciously using decoupling capacitors [42]. The specific challenges in power distribution for the Hero module result from the large number of chips placed close to each other, the need to supply cooling and power to alternate locations on each face of the array, and the need to exert no excessive forces on the chips which would disrupt their alignment. For example, placing power conversion circuitry close to the chips complicates the cold plate, which must cool one layer of chips while providing the required clearance to the power components serving the other chip layer.

5. On-chip

Machines built from tiled VLSI chips using Proximity Communication will need to support high bandwidth across those VLSI chips. Each chip would not only have a compute or memory core but also part of a distributed network that physically spans the entire system, routing packets internally on each chip and hopping between chips on Proximity Communication.

This network uses standard VLSI wires and needs high bandwidth, acceptable latency, good reliability, and low power. The first three constraints are easily met. Bandwidth to match the Proximity Communication links

between chips comes from the high density of on-chip wires. Wire latency, hampered by the resistive nature of on-chip wires, can be minimized through wire engineering and repeater insertion. Reliability arises from circuits such as differential signaling, low-offset receivers, and asynchronous handshaking [43][44].

However, the low power constraint is more difficult. A VLSI wire has a total capacitance well approximated by four parallel plate capacitors plus a fringe term [46],[45]; for typical wires, side-to-side capacitance dominates the total loading. Wire cap can be approximated at 0.3pF per mm of length ignoring switching effects, and this value changes little under technology scaling. A wire spanning a 15mm die, then, has a total capacitance of 4.5pF, and 256 64-bit buses criss-crossing the chip as part of a distributed system network represent a load of 75nF. The power required to swing this capacitance through the power supply is $P \propto C_{total} V_{dd}^2 f$. At a clock frequency of 4GHz, a power supply of 1V, and an activity factor of perhaps 0.1, we end up consuming nearly 30W just in network activity. Add in the extra capacitance of repeaters and our wires approach nearly 50W of dissipated power. Clearly, we need some form of power reduction for our on-chip wires.

Many schemes of efficient VLSI wiring have been built [47][48]. They dramatically lower power by reducing signal swing. By not reducing the power supply as well, these schemes return a linear power savings but avoid the complexity of generating and distributing multiple power supplies to a high-performance VLSI chip. They also employ amplifiers at the receivers to magnify small swings back to full CMOS voltage levels.

However, low-swing circuits have a couple of important limitations. First, pushing a reduced signal swing step onto a long wire still needs a large driver due to the wire's dominant RC time constant, and these large drivers end up consuming most of the power in a low voltage-swing system. Second, the on-chip wires are diffusive, so successive symbols blend and interfere with each other, reducing the fidelity of data transmission and ultimately limiting performance. Channel pre-equalization techniques common to board-level signaling are too expensive for the multitude of on-chip drivers.

We propose a wire system that overcomes these difficulties. It provides a small-swing signal to on-chip wires by using capacitive dividers created by simply spacing on-chip wires close together. This exploits what is usually seen as a drawback: the fact that side-to-side capacitance dominates total capacitance. Figure 9 shows an example circuit schematically. A driver drives the "pitchfork" structure that capacitively couples to the long wire through facing "tines" of the pitchfork by a total summed capacitance of C_c.

In this scheme, wire drivers can be tiny, because the capacitance they see is reduced: the small coupling capacitor is in series with the big wire capacitance. This lowers the power, area, and complexity of the circuit. The wires swing at a voltage of $\frac{C_c}{C_{total}}$, without the need for any reduced power supplies or step-down circuitry. Also, the capacitors offer a significant pre-emphasis signal boost. Because capacitors look like short-circuits to high-frequency signals, the edge of a signal transition is transmitted significantly faster, reducing the interference otherwise present between successive signals, and hence improving performance. Repeaters using this circuit can be placed back-to-back across a chip, and rolled into logic such as routing or error correcting.

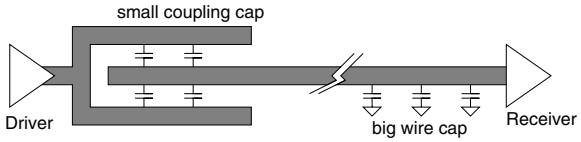

Figure 9. Wires driven by coupling capacitors

These circuits have some complications. The long wire needs to be appropriately biased, because it is AC-coupled to the driver. Also, the voltage swing is fixed in fabrication, so any dynamic swing adjustments require extra circuitry. However, because the coupling capacitor is made of the same structures composing the wires, variations in the wires should affect both total capacitance and coupling capacitance, minimizing process-induced variations. The receiver, as with any low-swing system, requires a low-offset sense amplifier to restore signals to full voltage levels.

We have built silicon implementing these designs in our lab and have found the ideas promising for reducing power while providing the bandwidth, latency, and reliability required for high-bandwidth systems.

6. Optical module-to-module

Parallel optical interconnects have had significant penetration in box-to-box interconnect applications. Previously, optical interconnects could provide system bandwidth on the order of a few Tb/s: enough to enable interconnection of modest arrays of processor/memory units. The Hero platform requires modules with a significant increase in optical interconnection capacity over existing systems. This will necessitate vast numbers of optical modules and fiber optic cables. Figure 10 illustrates a notional depiction of a Hero multi-chip module which contains multiple processor and memory chips and an interconnect sub-module with a capacity in excess of 100Tb/s. Providing this bandwidth brings some major challenges:

- Reliability
- Scalable physical transport medium
- Integration
- Cost

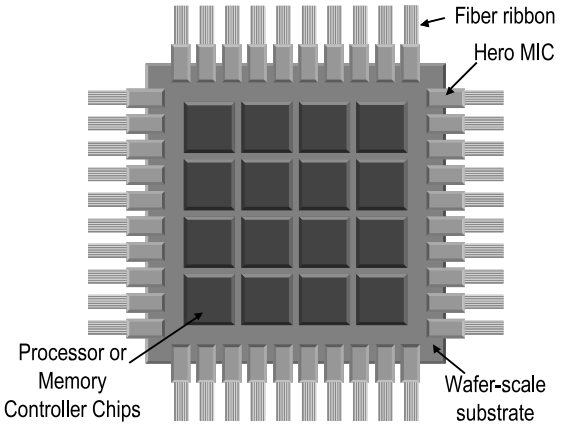

Figure 10. Hero module with Module Interconnect Chips (MICs)

In a high-productivity environment, reliability and availability are critical performance parameters. A typical metric is "five 9's availability," which means the server should be available 99.999% of the time (or have less than 4 minutes of down time per year). If we optimistically assume 2 FITs (failure-in-time per billion device hours) per single vertical-cavity surface-emitting laser (VCSEL) and 1,000,000 VCSELs per system, then the total FIT of the optically interconnected system would be 2,000,000. This translates to 1 failure per 500 hours and falls far short of the five 9's requirement. Therefore, reliability improvements, redundancy and innovative device-level solutions are required. Fortunately, strong progress in VCSEL reliability and recent innovations in high-density optical modules are improving system availability.

The physical transport medium for Hero must be scaleable to multiple Tb/s per chip. Fiber and polymer wave-guides are obvious candidates. However, neither presently supports the sheer number of connections at the interconnect distance and density required by the modern processor chip. Free-space interconnects have been proposed, but a platform consistent with mainstream manufacturing flow or thermal constraints has not yet emerged. Again, for an intimate integration, the number (and associated cost) of fiber connectors must be reduced. Short reach interconnects to processors based on Wavelength-Division Multiplexing (WDM) have been previously suggested [49], and appear to be a necessity for the Hero platform.

For the foreseeable future, electrical VLSI circuits will be responsible for processing information. Therefore, any optically interconnected system will involve optical-to-electrical conversion and vice-versa. Delivering data to the optical components and breaking electrical bottlenecks becomes one of the most critical issues for optical transceivers, particularly if the photonics and electronics are not tightly integrated. The first electrical bottleneck appears between the optoelectronic driver and receiver circuits and the photonic devices themselves. One way to resolve this bottleneck would be to tightly integrate the optics and electronics. This can be addressed by integrating optical devices directly onto the silicon circuits using, for instance, flip-chip attachment [50][51]. A second electrical bottleneck appears in the data transport from the processor to the input of the optical transceiver. Here, there are no conventional electrical interconnect solutions for off-chip bandwidth beyond 2 Tb/s. This bottleneck is more difficult to solve due to a lack of very high-speed bus standards among processor vendors, and additionally because there is no credible packaging or thermal solution for an integrated processor-plus-photonics offering. Hero solves this issue by using Proximity Communication, and co-locating high-density Proximity and optical transceiver circuits.

The final issue for penetration on a massive scale is cost, which is highly dependent on target volumes and technology investment. Present cost curves stem from a low integration level of optical transceivers and a low production volume. As evidenced by the semiconductor industry over the past several decades, a higher integration level enables lower-cost production. In addition, packaging optical chips into the transceiver and the resulting testing at the various stages incurs a significant fraction of the overall cost of the optics. Hence we may simultaneously solve integration and cost issues with a tightly integrated optical transceiver chip that removes both electrical and optical data transport bottlenecks.

For these reasons, we are investigating an integrated optical Module Interconnect Chip (MIC) technology based on Wavelength-Division Multiplexing (WDM), as shown in Figure 11. Each MIC provides optical interconnect with an I/O capacity in excess of 2 Tb/s. Additionally, we intend to incorporate into these MICs a proprietary, high-speed electrical interconnect based on Proximity Communication to route electrical information from silicon-based processors, memory, and I/O control chips to possible non-silicon photonic MICs. The high density of Proximity Communication enables the combination of diverse silicon ASICs onto a common platform with several orders of magnitude increase in off-chip communication capacity. This enables heterogeneous multi-module systems to be designed with a seamless communication capacity across the system, that removes a hierarchy of capacity-limited communication structures.

7. Sea-of-Anything (SOX) Interconnect

The architecture of the Hero system is predicated upon massive parallelism in a single shared memory address space. This design requires all resources in the system, including processors, memory controllers and IO buses, to individually interact with each other. However, the physical connectivity constraints of the Proximity and optical

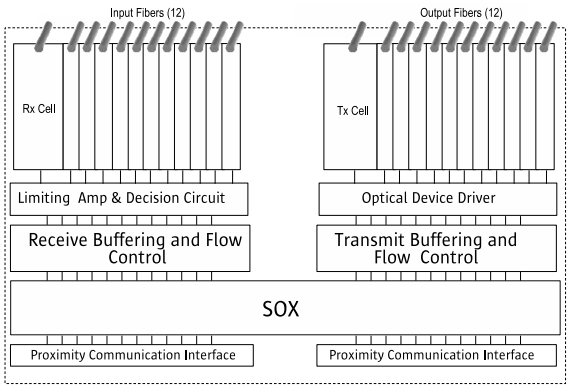

Figure 11. Components of a MIC

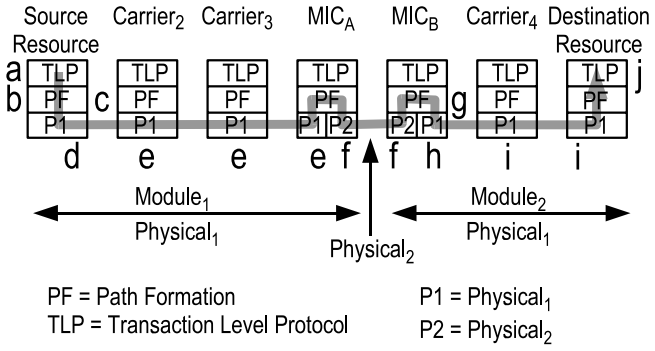

Figure 12. Sending a bus cycle transaction through the SoX interconnect

communication technologies enforce a strict topology on Hero: a set of Manhattan meshes built using Proximity Communication interconnected with point-to-point optical links. This, in turn, requires a transport mechanism that transparently enables any resource to transmit data to any other resource. SoX is a distributed switching transport interconnect that satisfies this requirement and maintains low latency, while taking full advantage of the underlying high-bandwidth physical communication technologies. SoX minimizes latency by employing cut-through switching [55] and worm-hole routing [56]. Furthermore, SoX minimizes latency by maintaining a consistent packet format as it traverses multiple physical communication technologies. That is, there is no adaptation processing required when a packet moves from one physical communication technology to another.

Since SoX is designed to operate in a system that supports a single address space, each resource in the system is assigned a destination address range within the single system address space. When a source wishes to communicate with another part of the system, the destination system address uniquely identifies the destination.

Below this global single address space, SoX implements a distributed worm-hole switching system that moves packetized bus cycles from the source to the destination. The path from the source to the destination transparently traverses many Proximity Communication hops and optical communication links. The source examines the destination address in a packetized bus cycle and converts the address into a path sequence of lower-level switching commands. Packets traverse along this defined path through Proximity Communication meshes and the optical communication links by sequentially processing the simple switching commands encoded in the path, similar to Myrinet [52].

Above the distributed switching layer, SoX supports a set of transactional protocols implementing memory coherency semantics. This transactional layer defines interactions between source and destination resources for two-party transactions, and between source, destination and forwarded-destination resources for three-party transactions. The transactional layer protocols are uniform throughout the entire system, independent of underlying physical layers.

Figure 12 illustrates the sequence of actions that take place for a SoX resource to send a transaction to a destination resource on another module in the Hero system. In this example, the source on $Module_1$ wants to send a packet to a destination on $Module_2$. The source first determines a path to the destination resource. This path goes through MIC_A, via $carrier_2$ and $carrier_3$, and so the source prepends this path segment to the packet. Upon arriving at MIC_A, the packet traverses a link to MIC_B which prepends to the packet another path segment to the destination via $carrier_4$. The sequence of steps and their corresponding location on Figure 12 are:

(a) The source packetizes the bus cycle transaction.
(b) The source determines the location of the destination resource and the path to it.
(c) The source appends a sequence of switching commands to the packet. The commands form a path from the source to MIC_B.
(d) The packet enters the SoX distributed switching meshes at the source.
(e) The packet moves from resource to resource until it gets to MIC_A.
(f) The packet traverses the optical communication link from MIC_A to MIC_B.
(g) Upon arriving at MIC_B, another path segment is appended to the packet that contains the switching command sequence to get to the destination.
(h-i) The packet moves from resource to resource until it gets to the destination.
(j) When the packet arrives at its destination, it is checked for data integrity, translated back into a bus cycle transaction and processed.

Due to the massive scale of the Hero system, SoX poses architecture and implementation challenges, including:

- Avoiding deadlocks in distributed switching meshes
- Achieving fault tolerance as parts of the distributed switching fabric fail
- Maintaining and updating the path tables that resources use to form switching command sequences when creating a packetized bus cycle

The combination of worm-hole routing and Manhattan switching meshes creates the challenge of avoiding deadlocks in the switching meshes. Packet switching deadlock occurs when packets block each other in a cyclical manner. SoX avoids the problem of packet switching deadlock by guaranteeing that the union of all possible packet paths conforms to a Directed Acyclic Graph [57], regardless of how malformed packets may become, thus eliminating the possibility for cyclic dependencies [58, 59].

Addressing the fault-tolerance challenge requires detecting faults which may be caused by failures of resources or parts of the communication fabric, and taking immediate corrective actions. These actions may vary from a simple retransmission to reassigning communication and computing tasks to alternate resources. One of the techniques employed to detect faults is to require all SoX transactions to be positively acknowledged. So when a source does not receive a positive acknowledgement within a certain timeout period, a potential fault has been detected. This timeout event triggers an immediate retransmission of the transaction on an independent, disjoint path. Continued timeouts initiate an out-of-band system to locate the source of the problem and prepare a new set of paths to isolate and circumvent the faulty resources or communication links.

Addressing the path table challenge requires the ability for all path tables in a SoX interconnect to be recomputed and updated upon any change of a system's topology. Some of the events that trigger this activity include, but are not limited to, system power-up, isolation of faulty components or the re-integration of repaired components. When most resources are operational in a system, the computation of path tables is relatively simple. However, when a system has many components that have been isolated due to failure, the updating of these path tables becomes much more complicated. Research into algorithms that can cope with the scale of a Hero class system and handle large numbers of random failures is ongoing.

8. Summary

This paper examines the memory bottleneck in large-scale computers, exploring it from both physical and functional viewpoints and considering system balance using the ratio of $bytes/flop$ versus memory capacity. Additionally, it proposes a computer we term "Hero," that aims to achieve a much flatter memory hierarchy.

The paper presents the key features and challenges of three high-bandwidth physical technologies that can be used in Hero: Proximity Communication, low-power on-chip networks, and dense optical communication. Proximity Communication provides orders of magnitude improvement in off-chip bandwidth compared to traditional I/O, but has key alignment challenges. We develop a simple analytic model for the trade-off between alignment accuracy and achievable bandwidth. Our on-chip networks provide ample bandwidth and address power concerns with a new capacitively-coupled driver circuit. Optical communication uses dense WDM combined with Proximity circuits to address important optical challenges.

In addition, the paper discusses a transport layer, SoX, that is a provably deadlock-free high-bandwidth low-latency two-level interconnect fabric. SoX uses Proximity Communication meshes for inter-chip communication and WDM fiber optics for inter-module communication. SoX also includes means to detect and to handle deadlocks that arise from equipment malfunction.

Together these technologies have the potential to flatten the bandwidth hierarchy of large-scale computers to maintain high $bytes/flop$ across all levels of caching, out to the full memory footprint of a machine. Presently, most code is planned and generated with the expected constraint of memory hierarchies that provide ever-decreasing bandwidths to larger sets of memory. With the technologies described here, hardware and software architects can look towards designing systems, compilers, and applications that effectively utilize a much flatter bandwidth hierarchy.

9. Acknowledgments

We acknowledge every member of Sun's HPCS team, led by Jim Mitchell and supported by Greg Papadopoulos. In addition we recognize the effective support and guidance from DARPA as part of its HPCS Phase II program.

References

[1] D. Pham, *et al.*, "The Design and Implementation of a First-Generation CELL Processor," *IEEE ISSCC*, Feb 2005.
[2] W. Beyene *et al.*, "Design and analysis methodologies of a 6.4 Gb/s memory interconnect system," *Electronic Comp. and Tech.*, pp. 1406-1411, June 2004.
[3] "XDR XIO Data Sheet Summary," *www.rambus.com*.
[4] "PCI-X 2.0: High Performance, Backward Compatible PCI for the Future," *www.pcisig.com*.
[5] R. Ho, K. Mai, and M. Horowitz, "The Future of Wires," *Proceedings of the IEEE*, pp. 490-504, April 2001.
[6] *Int'l Technology Roadmap for Semiconductors*, 2003.
[7] C.-K. K. Yang *et al.*, "A 0.5um CMOS 4Gb/s serial link transceiver with data recovery using oversampling," *IEEE Journal of Solid-State Circuits*, pp. 713-722, May 1998.
[8] R. Drost and B. Wooley, "An 8-Gb/s/pin Simultaneously Bidirectional Transceiver in 0.35-micron CMOS," *IEEE Journal of Solid-State Circuits*, pp. 1894-1908, Nov. 2004.

[9] J. Gustafson, "Purpose-Based Benchmarks" *Int'l J of High Perf. Computing Applications*, Vol. 18, pp. 475-487, 2004.
[10] J. McCalpin, "Memory Bandwidth and Machine Balance in Current High Performance Computers," *IEEE Technical Committee on Comp. Arch. (TCCA) Newsletter*, Dec 1995.
[11] T. Chilimbi *et al.*, "Making Pointer-Based Data Structures Cache Conscious," *IEEE Computer*, pp. 67-74, Dec 2000.
[12] H. Meuer *et al.*, Universities of Mannheim and Tennessee, *Top500 Supercomputer Sites*, www.top500.org.
[13] N. Adiga *et al.*, "An overview of the BlueGene/L supercomputer," *Proc. ACM/IEEE Conf. Supercomputing*, 2002.
[14] F. Allen *et al.*, "Blue Gene," *IBM System Journal*, Vol. 40, No. 2, 2001, pp. 310-327.
[15] D. J. Kerbyson *et al.*, "A comparison between the Earth Simiulator and AlphaServer systems using predictive application performance models," *Proc. Intl. Parallel and Distributed Processing Symp.*, April 2003.
[16] *SGI Altix 3000 Hardware*, www.nas.nasa.gov/Users/ Documentation/ Altix/hardware.html.
[17] T. Dunigan *et al.*, "Early Evaluation of the Cray X1," in *Proc. ACM/IEEE Conf. Supercomputing*, 2003.
[18] H. Shan and E. Strohmaier, "Performance characteristics of the Cray X1 and their implications for application perf. tuning," *Supercomputing Conf.*, pp. 175-83, 2004.
[19] K. Krewell, "UltraSPARC IV mirrors predecessor," *Microprocessor Report*, November 2003, pp. 1-3.
[20] *Sun Fire E25K Datasheet*, www.sun.com/ servers/highend.
[21] A. J. van der Steen and J. Dongarra, "Overview of recent supercomputers," www.top500.org/ORSC/2004.
[22] *Ultra2 Workstation Architecture*, Sun Microsystems, 1999.
[23] T. Sato, *et al.*, "Earth Simulator Running," *Intl. Supercomputing Conf.*, Heidelberg, June 2002.
[24] S. Habata, *et al.*, "The Earth Simulator system," *NEC Research and Development*, Vol. 44, January 2003.
[25] *UltraSPARC IV Processor Architecture Overview*, Sun Microsystems technical white paper, 2004.
[26] S. Naffziger, *et al.*, "The implementation of the Itanium 2 microprocessor," *IEEE Journal of Solid-State Circuits*, Vol. 37, No. 11, November 2002, pp. 1448-1460.
[27] H. Sharangpani and K. Arora, "Itanium processor microarchitecture," *IEEE Micro*, pp. 24-43, Sept-Oct 2000.
[28] R. E. Kessler, "The Alpha 21264 Microprocessor," *IEEE Micro*, Vol. 19, No. 2, March-April 1999, pp. 24-36.
[29] J. Borkenhagen *et al.*, "A multithreaded PowerPC processor for commercial servers," *IBM Journal of Research and Development*, Vol. 44, No. 6, November 2000, pp. 885-898.
[30] A. Klauser, "Trends in high-performance microprocessor design," *Telematik-2001*, No. 1, pp. 12-21, 2001.
[31] R. E. Kessler, *et al.*, "The Alpha 21264 Microprocessor Architecture," *IEEE ICCD*, pp. 90-5, Oct 1998.
[32] B. Khailany *et al.*, "Imagine: media processing with streams," *IEEE Micro*, Vol. 21, pp. 35-46, Mar-Apr 2001.
[33] U. Kapasi *et al.*, "The Imagine stream processor," *Proc. 2002 IEEE Intl. Conf. Computer Design: VLSI in Computers and Processors*, pp. 282-288, September 2002.
[34] B. Serebrin *et al.*, "A stream processor development platform," *IEEE ICCD*, pp. 303-8, Sept 2002.
[35] L. Oliker *et al.*, "Scientific computations on modern parallel vector systems," *Proc. ACM/IEEE SC2004 Conf. Supercomputing*, November 2004, pp. 10.
[36] R. Drost, *et al.*, "Proximity Communication," *IEEE JSSC*, pp. 1529-35, Sept. 2004.
[37] R. Drost, *et al.*, "Electronic Alignment for Proximity Communication," *IEEE ISSCC*, Feb. 2004.
[38] M.-J. E. Lee *et al.*, "Low-Power Area-Efficient High-Speed I/O Circuit Techniques," *IEEE JSSC*, Nov. 2000.
[39] W. J. Shakespeare *et al.*, "MEMS Integrated Submount Alignment for Optoelectronics," *Journal Of Lightwave Technology*, pp. 504-509, 2005.
[40] D. W. Copeland *et al.*, "Manifold Microchannel Heat Sinks: Isothermial Analysis," *Proceedings, InterSociety Conference on Thermal Phenomena*, pp, 251-257, 1966.
[41] B. Guenin, "Calculations for Thermal Interface Materials," Electronics Cooling, Vol 9, No. 3, August, 2003.
[42] L.D. Smith, *et al.*, "Power Distribution Sys. Design Meth. and Capacitor Selection for Modern CMOS Tech.," *IEEE Transactions on Advanced Packaging*, pp. 284-91, 1999.
[43] I. Sutherland, J. Lexau, "Designing Fast Asynchronous Circuits," *IEEE Async 2001*, pp. 184-193, March 2001.
[44] R. Ho, J. Gainsley, R. Drost, "Long wires and asynchronous control," *IEEE Async 2004*, pp. 240-249, April 2004.
[45] R. Ho, *et al.*, "Managing Wire Scaling: A Circuit Perspective," *IEEE Int'l Interconnect Tech. Conf.*, Jun 2003.
[46] M. Bohr, "Interconnect Scaling–The Real Limiter To High-Performance ULSI," *IEEE Electron Devices Meeting*, pp. 241-244, December 1995.
[47] H. Zhang *et al.*, "Low-swing on-chip signaling techniques," *IEEE Transactions on VLSI*, pp. 414-419, April 1993.
[48] R. Ho, *et al.*, "Efficient On-Chip Global Interconnects," *IEEE Symposium on VLSI Circuits*, June 2003.
[49] A. V. Krishnamoorthy *et al.*, "AMOEBA: an optoelectronic switch for multiproc. networking using dense-WDM," *IEEE J. Special Topics in Quantum Elec.*, pp. 261-75, Mar 1999.
[50] K. W. Goossen and A. V. Krishnamoorthy, "Optoelectronics-in-VLSI," in *Wiley Encyclopedia of Electrical and Electronic Engineering Vol. 15*, pp. 380-395, 1999.
[51] A. V. Krishnamoorthy *et al.*, "Fibre-to-the-chip: VCSEL arrays for integration with VLSI circuits," in *Handbook of Laser Tech. and Applications: Vol. 3 Ed. C.*, UK, Dec. 2003.
[52] N. Boden, *et al.*, "Myrinet: A Gigabit-per-second Local Area Network," *IEEE-Micro*, pp. 29-36, Feb 1995.
[53] P. Palnati *et al.*, "Deadlock-free routing in an optical interconnect for high-speed wormhole routing networks," *ICPADS*, 1996.
[54] J.E. Cunningham, *et al.*, "Scaling VCSEL reliability up to 250Terabits/s of system bandwidth," *OSA Topical Meeting on Information Photonics*, June 2005.
[55] P. Kermani and L. Kleinrock, "Virtual Cut-through: A New Computer Communication Switching Technique," *Computer Networks*, pp. 267-86, Sep. 1979.
[56] W. J. Dally, and C. L. Seitz, "Deadlock-free message routing in multiprocessor interconnection networks," *IEEE Transactions Comp.*, pp. 547-53, May 1987.
[57] F. Harary and E. M. Palmer, *Graphical Enumeration*, Academic Press, New York, 1973.
[58] C. Glass and L. Ni, "The turn model for adaptive routing," *J. ACM*, pp. 874-902, 1994.
[59] E. Fleury *et al.*, "A General Theory for Deadlock Avoidance in Wormhole-Routed Networks," *IEEE Trans. Parallel Distrib. Syst.*, pp. 626-38, 1998.
[60] J.-S. Yang and C.-T. King, "Designing Deadlock-Free Turn-Restricted Routing Algorithms for Irregular Wormhole-Routed Network", *J. Information Science and Engineering*, pp. 575-94, 2001.