White Paper

Competing Against IA-32 Clusters

Connie Waring and Rex Tanakit

SGI Proprietary and Confidential
For Internal Use Only

Competing Against IA-32 Clusters

Summary

There are applications and situations where IA-32 clusters will perform very well. We need to recognize these circumstances and respond accordingly. This white paper is intended to give you pointers on when to compete and when to walk away. This paper targets systems engineers and sales representatives. ***NOTE: This is an SGI internal document ONLY.***

1.0 What Is a Cluster?

A cluster is a collection of single-system image (SSI) machines that are coupled together through some form of communications network, software protocols, and tools that function as a single entity for certain applications. The individual machines can be of any type and any size. In the past SGI has promoted the idea of clustered machines. In fact, in the mid-1990s, one of the successful product offerings from SGI was the Power ChallengeArray™ (see Figure 1).

Typically a cluster with an IA-32 Intel® architecture is configured as a "Beowulf"-style cluster (see Figure 2)- that is, it is constructed of low-cost commodity parts and open-source software with the goal of creating a high-performance or supercomputer-class machine from this distributed processing environment[1]. Usually, the cluster is composed of compute nodes, but there can be additional special-purpose nodes, such as the following:

Fileserver nodes

- Head nodes (the front-end node that users access and submit jobs)
- Graphics nodes for visualizing output
- Other special-purpose nodes

Although most IA-32 clusters run some flavor of Linux®, it is becoming much more common to see clusters running the Microsoft® Windows® operating systems, usually Windows NT® or Windows® 2000.

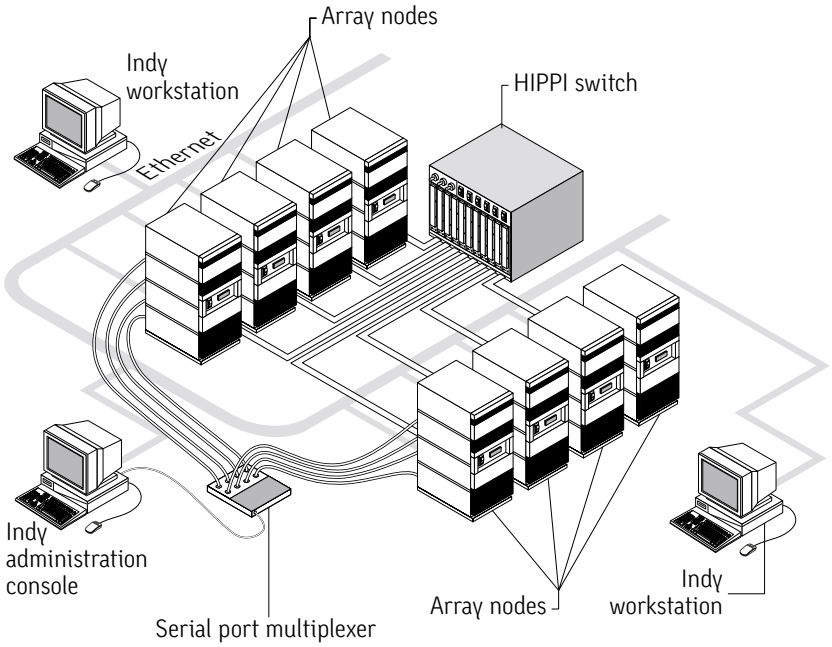

Fig. 1. Power ChallengeArray, an early clustered offering from SGI

1. Beowulf *Project-www.beowulf.org*. Donald Becker, CESDIS Project, NASA's Goddard Space Flight Center.

SGI Proprietary and Confidential
For Internal Use Only

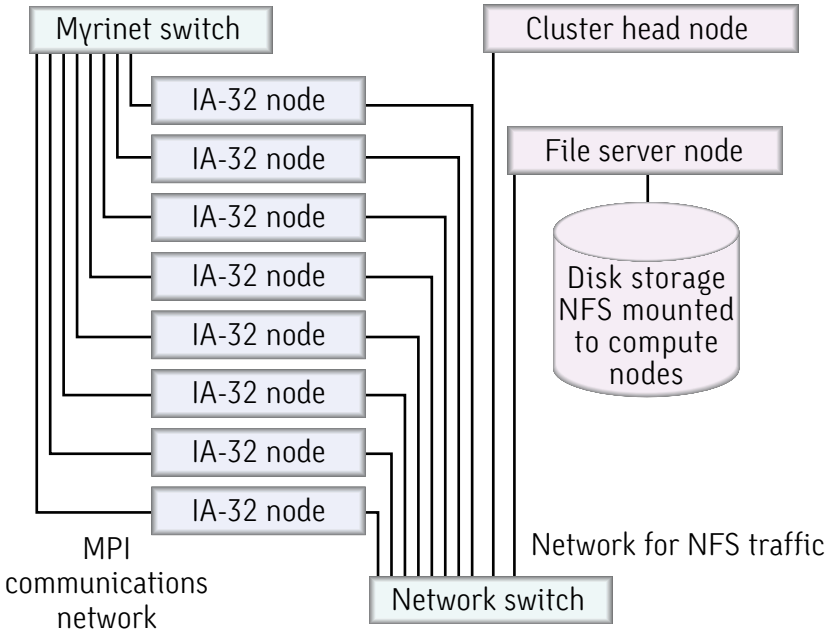

Fig. 2. Typical IA-32 Beowulf-style cluster

A typical quote for a Beowulf-style IA-32 cluster might be similar to the following (taken from an actual customer's quote). Note that the high-speed interconnect for message passing constitutes almost 35% of the total cost.

Quantity	Description	Cost (ea)	Total Cost
32	Rackmount 2U chassis, Intel D850MVSE montherboard, Intel Pentium® 4 1.8 GHz Ssocket 478 .13 micron CPU, 1024GB PC800 RDRAM, 60GB disk, CD-ROM, floppy, Red Hat Linux 7.2	$2,050.00	$65,600.00
32	Dolphin Interconnect SCI 2D torus, 64-bit/66 MHz, 1 slot, all required cabling included	$1,615.00	$51,680.00
32	Scali ClusterEdge Management Software	$600.00	$19,200.00
1	HP Procurve 4000M Ethernet switch chassis, 48 ports initialized	$2,315.00	$2,315.00
3	Trimm Technologies 40U equipment cabinet with shipping crate	$2,350.00	$7,050.00
1	48-port Cyclades rackmount terminal server	$4,320.00	$4,320.00
2	Annual maintenance support contract for hardware and software	$7,868.00	$15,736.00
		Total	**$165,901.00**

It is very important that the customer compares all of the system's components when comparing IA-32 cluster with SGI servers. The SGI Origin family servers come with the built-in NUMAlink interconnect. Furthermore, the NUMAlink™ has a lower latency and higher bandwidth than the commodity-off-the-shelf interconnect. The advantages result in a better scalability of the application when running in parallel mode.

2.0 SMP with SSI Versus Cluster Architecture

Symmetric multiprocessor (SMP) architecture with an SSI offers a degree of flexibility and usage of the machine that is unavailable with clusters. SMP allows for a much more general-purpose machine. Also, it can run efficiently programs targeted for most parallel programming architectures.

SGI Proprietary and Confidential
For Internal Use Only

One of the key advantages of an SMP architecture machine is that the system has very fast access to a very large reserve of common memory. By contrast, the amount of memory available to each CPU within an IA-32 cluster is very limited.

In addition, there are associated total cost of ownership (TCO) factors that may be less with large SSI machines, such as reduced system administration, reduced data center space cost, lower power consumption, lower setup cost, and lower upgrade costs in the future. For more internal SGI information on TCO, see: *http://chempharm.engr.sgi.com/htc_bioinfo/011025_tco_doc.html* or *https://channel.sgi.com/res/default/solutions/sciences/index.html*.

3.0 SGI Machines Can Be Used As Clusters Also!

SGI has been successful with many different models of clustered machines. In 1994 SGI demonstrated running seismic processing codes across an array of 10 Power Challenge™ servers, each with 18 R8000 CPUs.

In 1996 SGI used an array of four Power Challenge servers, each with 16 R8000 CPUs, to run a bioinformatics code that analyzed the entire yeast genome in a live Web event called GeneCrunch.

The 1998 ASCI project in Los Alamos provided an outstanding example of using SGI Origin family machines in a clustered configuration. This cluster comprised 48 SGI® 2000 series servers (with 128 processors each, for a total of 6,144 processors), 1,500+GB of main memory, and 70+TB of disk storage. The interconnect infrastructure was HIPPI based.

Other customers have been successful with clustering groups of SGI® 2400 servers and SGI® Origin® 200 servers. Currently, customers are purchasing SGI® Origin® 300 clusters for areas such as bioinformatics or crash simulation.

Any of the SGI Origin server family can be clustered using interconnects ranging from 10baseT to Myrinet™-2000 (*www.myri.com*) to Gigabit System Network to NUMAflex™ to create a very flexible cluster architecture, with a very mature operating system and tools to effectively utilize it.

In many cases the customer must construct an IA-32 cluster with hardware and software components from different vendors, each with their own support system. The customer could also work through a Linux support service, which functions as an intermediary between vendors but has no influence on them. SGI improves this situation by acting as a single-source vendor.

One of the advantages of an SGI cluster is that it is based on 64-bit SMP nodes, giving the flexibility to utilize the advantages of a 64-bit SMP-type system as well as the incremental growth and aggregate performance possibilities of a clustered environment.

4.0 Where IA-32 Clusters Are Successful

Applications that tend to be "embarrassingly parallel" and small enough to fit into a 32-bit address space, such as almost all seismic imaging codes (particularly prestack time migration and depth migration codes) and many bioinformatics codes, run very well on IA-32 clusters. An "embarrassingly parallel" application does very little communication between nodes. For example, in some seismic imaging codes very low communication is required except at the beginning (where certain data is broadcast) and at the end, where partial images are combined into the final image. The majority of time is spent in very intensive number crunching and local disk I/O. A new piece of data is occasionally obtained from the network, either via File Server, Network File SystemNFS or from the master. In computational fluid dynamics (CFD) and crash simulation, when all phases of the applications fit within a 32-bit address space, IA-32 clusters can perform very well.

Another example of an application that is well suited to an IA-32 cluster is an animation program that renders hundreds of similar images to generate film frames. In this example, no communication is needed between nodes.

5.0 Applications and Requirements That Are Inappropriate for IA-32 Clusters

On IA-32 clusters the memory requirements for an individual process must fit within the 3GB RAM limitation of 32-bit Linux. Many application codes require RAM greater than what is available on individual IA-32 cluster nodes. Not having a 64-bit OS and address space is one of the biggest weaknesses of IA-32 clusters. For example, grid generation and partitioning phases of most CFD applications require greater than 3GB of memory.

If there is a large amount of interprocess communication between nodes, the performance of individual cluster nodes must be greater than the SMP to compensate for the communication times. High performance interconnects for IA-32 tend to be expensive, and there can be difficulties in expanding to larger numbers of cluster nodes. Typically, the available interconnects for IA-32 clusters have higher latency and lower bandwidths compared with SMP-style architecture interconnects, such as NUMAflex.

SGI Proprietary and Confidential
For Internal Use Only

Examples of applications that are not appropriate to clusters based on IA-32 include:

- Current parallel reservoir simulators, which are not coded very well. The communication required every time-step limits scaling on systems without very fast interconnects.
- Applications requiring very large dynamic databases visible to all nodes, including some graphical applications such as some seismic codes.
- Applications with large recursive dynamic data structures. Here parallelizing for distributed memory and getting good performance would be quite onerous.
- Applications that require a lot of I/O, particularly if large file support is needed.
- Operating system support for "non-PC"-class tape drives such as IBM 3490E, Ampex DST, and others.

Additionally, not all Linux filesystems support file sizes greater than 2TB.

6.0 Where SGI Origin Family Servers Can Win Against IA-32 Clusters

Win with Performance:

- Customers' applications have not been validated on the interconnect hardware and implementation of Message Passing Interface (MPI). Does the customer application ship with the manufacturer's MPI? Different versions of MPI library will result in different performance. SGI Origin family servers are fully functional systems with interconnect hardware and MPI library.
- In some cases running large models in applications such as Fluent (CFD) will show limitations in scalability (and therefore, absolute performance) when run in a clustered environment as compared with SSI. This is due to the higher interhost latency and also to the limited pipeline capability of the cluster's interhost communication.
- When scaling is an important issue for the customer, the SGI Origin server family offers superior scaling even on MPI applications up to 512 CPUs with the SGI® Origin® 3000 series or excellent scaling on a cluster of SGI Origin 300 servers with Myrinet-2000 interconnect.

Win with IRIX® (Mature OS with 64-bit Address Space):

- In CFD, typically, there are three phases in solving a problem:
  - Preprocessing-mesh/grid generation and partitioning
  - Solver
  - Postprocessing and visualization

  Customers often focus only on the solver, which IA-32 can perform very well. The preprocessing is an important step that may require 64-bit address space that IA-32 cannot process. We need to point this out to the customers. If the problem needs 64-bit, this is where we can win.
- If the customer needs operating system reliability and capabilities that are beyond what can be obtained with IA-32 Linux (including ongoing operating support). Many IA-32 clusters run on Linux, an operating system still widely considered in the industry to be immature. A failure on the Linux operating system can result in increased overall administration costs, because a significant amount of time and money can be spent on optimizing a specific application to run effectively on Linux. IRIX offers many advantages that cannot be currently found in Linux, such as:
  - Superior I/O handling
  - Trusted IRIX™
  - Superior compilers and performance profiling tools, as well as overall system monitoring tools such as Performance Co-Pilot™
  - More robust support for peripherals, particularly tape drives
  - Checkpoint-restart capability (for long-running jobs)
  - Weightless processes (to maximize system utilization)
  - Large memory page sizes (to reduce performance degradation due to TLB misses)

Win with Flexibility:

- When customers are writing their own applications, SMP architecture does not necessarily require MPI to run a parallel mode. In general, effective parallelization is always hard. However, it is true that the required coding effort is smaller and maintenance is easier on an SMP machine than on an MPI machine. In general, it is easier to engineer applications on an SMP machine because the application can draw from common memory without having to utilize message passing.
- If customers have future plans to expand their environment, the difficulties in expanding interconnect networks may be a valid point to bring up.

Win with Superior Service:

- SGI has knowledgeable engineers and one of the best service organizations to assist customers in their applications. Buying systems from SGI gives customers access to support that they could not get from vendors of IA-32 clusters.

SGI Proprietary and Confidential
For Internal Use Only

Win with Total Cost of Ownership:

- When customers acknowledge and are willing to consider the total cost of ownership. Based on a study done by the SGI bioinformatics team, a cluster of three SGI Origin 300 servers, each with 12 500-MHz CPUs, would do the equivalent work of a cluster of 32 1-GHz Pentium III machines, but the total cost of ownership would be only 58% of the cost when considered over three years.

For example, an SGI Origin 300 cluster vs. an IA-32 Linux cluster capable of obtaining the same performance on the bioinformatics application, BLAST, has vastly different power consumption:

	SGI Origin 300 Cluster			IA-32 Linux Cluster		
		Energy cost/year	Energy cost/3 years		Energy cost/year	Energy cost/3 years
Average watts/CPU	30			82.5		
Watts/hour system	360			6,240		
KW/h/year system	3,153.6			23,126		
Primary power $0.15/kWh	$0.15	$473		$0.15	$3,469	
Secondary power 0.6 PP	0.60	$284		0.60	$2,081	
Total power cost/year		$757			$5,550	
Power consumption cost/ 3 years			$2,270			$16,651

Other Benefits:

- If customers require a highly available environment, it must be considered that most PCs have a mean-time-between-failure of around one year. If you have a cluster of 512 processors-256 2p nodes-then this translates to a failure approximately every other day. This raises the question of whether the customer's application is fault-tolerant enough to compensate for having to withdraw a node from the cluster and to repair or replace it.
- If the customer's application requires real-time processing. Real-time processing is unavailable on Linux or Windows, and IRIX has a great deal of capability in this area.
- When the differentiating features available for SGI Origin family servers-beyond the mature operating system and hardware features-can give the customer an advantage. For example, the clustered XFS™ (CXFS™) product, not available for IA-32 Linux, can give a single, buffer-coherent, high-performance view of the journaled XFS filesystems to all nodes in the cluster.
- Complementary Products and Services: Even if the customer is set on having an IA-32 compute cluster, SGI Origin family machines can provide high-performance NFS servers, tape servers, and other functions in a more robust and reliable manner than can IA-32 systems.

  SGI can also offer complementary high-performance visualization solutions for imaging and graphics.

7.0 Where SGI Origin Servers Do Not Compete As Well—When Do We Walk?

- If the initial hardware cost is the customer's primary concern and the application already runs well in an IA-32 cluster environment, then there is no point in proceeding. We cannot compete solely on hardware costs when going up against commodity IA-32 hardware.
- When customers are already happy with IA-32, don't need the advantages of a 64-bit architecture and operating system, have a lot of system administrators to take care of their systems, or don't care about system administration cost, etc. A good example of this would be a university environment where students provide the system administration power for very low (if any) cost.
- When the application already runs well and scales to the extent required. If the application requires a fast interconnect, then the customer will have to face the burden of setting up a fast communication fabric (significant additional costs) and endure additional headaches if they want to grow its size. One item to consider here, however, is that IA-32 clusters do not scale linearly. If customers have benchmarked on an eight-node IA-32 cluster, they cannot be assured that they will receive anywhere near 3x performance enhancement on a 24-node cluster.
- If customers are using a "store-bought" application that has already been ported for IA-32 clusters and they feel it performs sufficiently, then in most cases they will feel that the significantly lower cost of the hardware outweighs other considerations.

SGI Proprietary and Confidential
For Internal Use Only

8.0 Case Study: How SGI Origin 300 Beat IA-32 Clusters

Recently, a university evaluated both a "mini-cluster" of two 8p SGI Origin 300 servers and a 70p Beowulf-style cluster. They decided in favor of an SGI Origin 300 server solution for the following reasons:

- The SMP SGI Origin 300 server option offered a degree of flexibility and usage of the machine unavailable with the Beowulf cluster option
- Many of their codes required RAM greater than what was available on the Beowulf-style nodes
- Although the Beowulf cluster included 54 more processors than the SGI Origin 300 server option, their benchmarks clearly showed the SMP option as the overall winner.
- There were two key factors involved in the benchmark: superior CPU performance and superior I/O (disk) performance
- Operating system and software reliability with the systems based on IRIX
- No need to purchase additional computer room air conditioning. The SGI Origin 300 server solution would put out only 6,500 BTU/hour, whereas the Beowulf-style cluster would produce more than 21,600 BTU/hour.

Please note: Several of the PC manufacturers that make machines specifically for high-density clusters (four processors in a 1-U form factor) are utilizing laptop versions of the processor chips in order to reduce the heat generation and power consumption. Most often when you see charts of heat and electricity consumption of MIPS® processors vs. Intel processors, the "normal" desktop flavor of the chips is being compared.

- Maintenance costs. The three-year maintenance quoted cost for FullCare? for the SGI Origin 300 server solution was only 34% of the cost of the maintenance quoted for the Beowulf-style cluster option.

  Please note: This ratio of maintenance costs may be specific to this installation. In the United States, in many instances, big-name PC manufacturers such as Dell and Compaq are providing three-year, next-day parts replacement service, free with the hardware.

- Partnership with SGI. Based on their past experiences with SGI, they felt that the advantages-such as technical liaisons, priority access to new SGI software products, loaner equipment, and roadmap meetings with SGI software and hardware experts, as well as project publicity-provided a valuable contribution.

9.0 Quick Positioning Guide

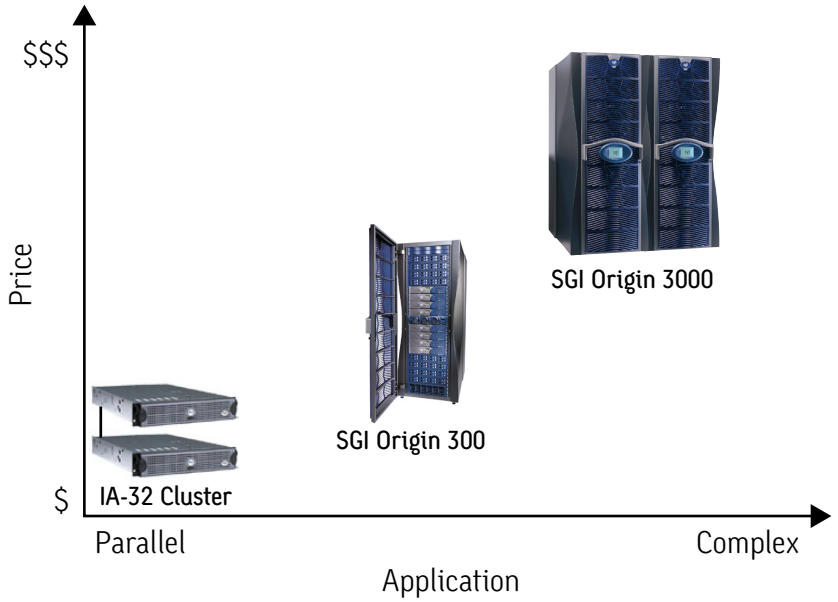

Fig. 3. Positioning SGI Origin Servers against IA-32 Clusters

SGI Proprietary and Confidential
For Internal Use Only

IA-32 Clusters:

- Low hardware costs
- High application programming complexity
- High overall system administration costs
- Additional cost for network infrastructure and management software
- Commodity off the shelf product

SGI Origin 300 Server:

- Mature/proven 64-bit OS
- SSI up to 32 processors in a compact form factor
- Scalable options for memory or storage

SGI Origin 3000 Series Servers:

- Mature 64-bit OS
- Scalability up to 512 processors in a single system image
- Highest performance system
- Ultimate flexibility
- High degrees of customization available

10.0 Acknowledgments

Many thanks to Haruna Cofer, Kumaran Kalyanasundaram, Brian Sumner, Mark Kremenetsky, and others for their help in the creation of this document.

Corporate Office
1600 Amphitheatre Pkwy.
Mountain View, CA 94043
(650) 960-1980
www.sgi.com

North America 1(800) 800-7441
Latin America (52) 5267-1387
Europe (44) 118.925.75.00
Japan (81) 3.5488.1811
Asia Pacific (65) 771.0290

© 2002 Silicon Graphics, Inc. All rights reserved. Specifications subject to change without notice. Silicon Graphics, SGI, Origin, IRIX, Trusted IRIX and the SGI logo are registered trademarks; Power Challenge, Power ChallengeArray, NUMAflex, NUMAlink, Performance Co-Pilot, XFS, and CXFS are trademarks; and FullCare is a service mark of Silicon Graphics, Inc. MIPS is a registered trademark of MIPS Technologies, Inc., used under license by Silicon Graphics, Inc. Intel and Pentium are registered trademarks of Intel Corporation. Linux is a registered trademark of Linus Torvalds. Microsoft, Windows, and Windows NT are registered trademarks of Microsoft Corporation. All other trademarks mentioned herein are the property of their respective owners
.

3269 (02/02) JI3627

SGI Proprietary and Confidential
For Internal Use Only