Cluster 2002

Blades—An Emerging System Design Model for Economic Delivery of High Performance Computing

Kirk Bresniker

Systems Architect
Internet and Applications Systems Laboratory

Hewlett-Packard

September 24, 2002

Cluster 2002

Agenda

Agenda

Introduction / Definitions
Origins of the Blade Server
Limitations of First Generation Designs
Second Generation Design
Future Directions

Cluster2k2_092402.ppt
Sept 24, 2002
Page 2

During the recent rise and subsequent fall of the internet bubble, a new computer system design model emerged, primarily from venture capital start-ups. Bladed Systems, dense arrays of single board computers housed in a common chassis, seemed a promising way for service providers to keep pace with the anticipated dot.com inspired build-out. The blades were dense, low bandwidth and low in computational power, but they were suited to rapid deployment of massive content delivery.

Along with other lessons learned as the irrational exuberance faded and unviable business models and their edge applications were winnowed from data centers, the designers of bladed systems began to realize that blades had the potential to move from edge-only applications into high performance enterprise, communication, and technical compute.

This discussion will cover the origins of the bladed server system model, the design, and limitations of the first generation designs, the ongoing technology changes which are enablers to second generation blades, their application to a wider set of multi-system compute problems, and some possible directions for future development.

Cluster 2002

Definitions

A Blade is …

An inclusive computing system that includes processor, memory,network connections and associated electronics on a single motherboard.

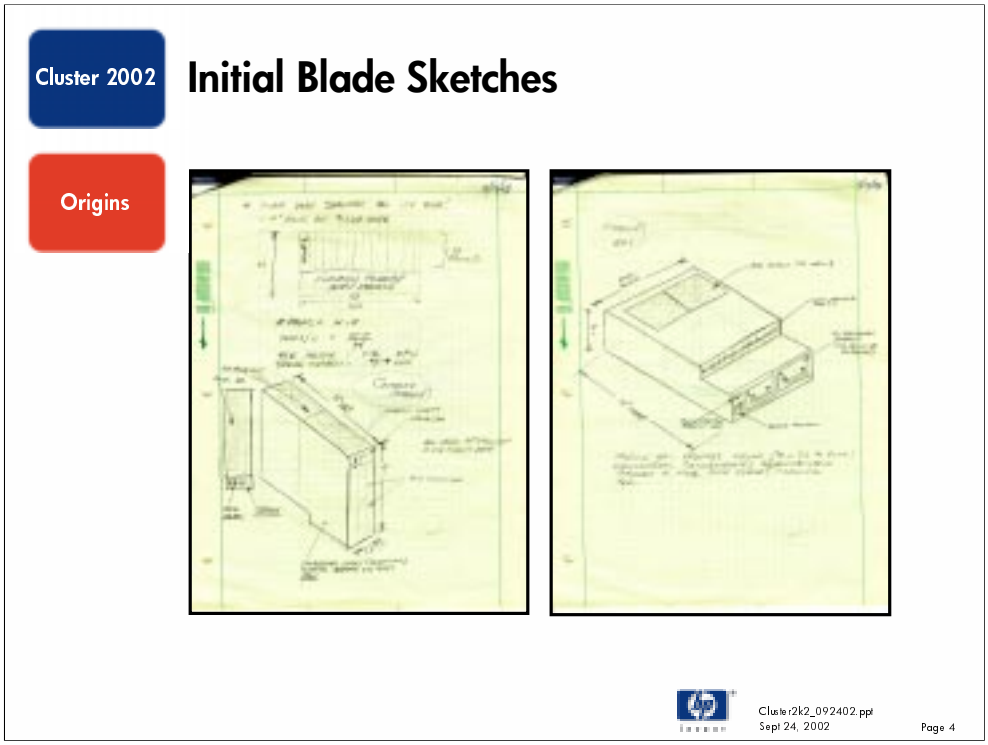
The server blade typically is associated with an enclosure system that allows multiple blades to be housed in a standard server 'sub-rack' or enclosure that share resources such as power supplies and cooling fans.

Blades …

- are easily accessible
- offer increased computing density
- have modular architecture that ensures flexibility and scalability.

Source: Provisioning the Internet Infrastructure: Server Blades and Dynamic Workload Management IDC Document # 24155

First, some definitions. According to IDC, A Blade is …



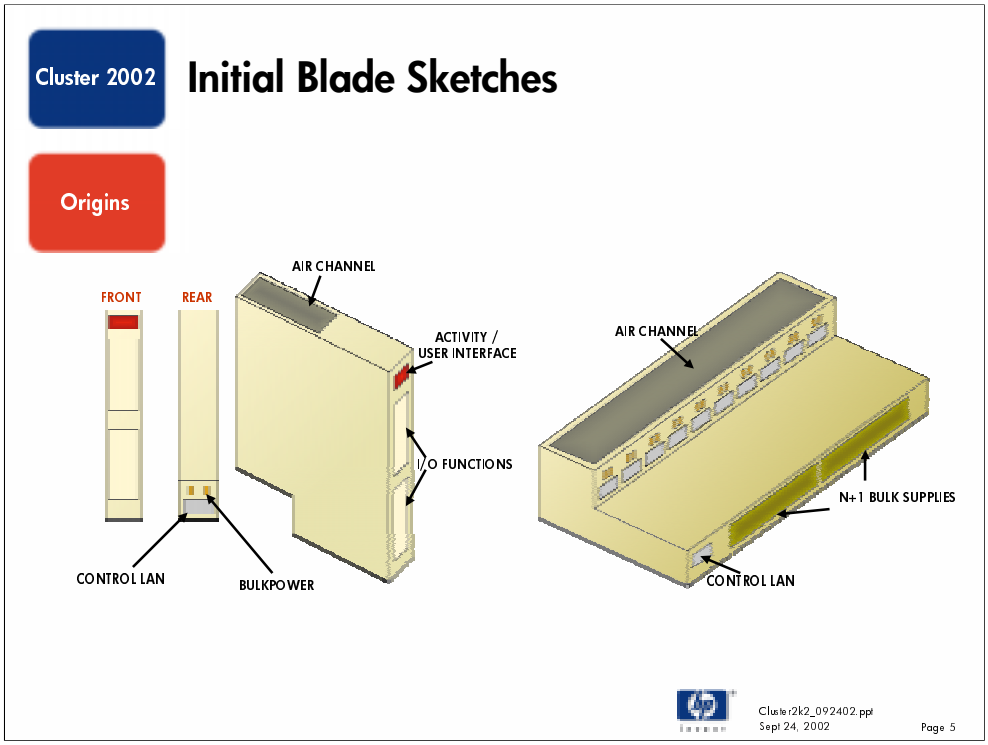
It is, perhaps, instructive is to dig back to the origins of the systems packaging concept that has come to fit this definition as it developed as a concept within Hewlett-Packard.

In late 1998, I went on a customer visit to the primary data center of a national internet service provider in Fairfax County, Virginia.

I had just completed architecting the first 2U HP-UX PA-RISC rack optimized server, after having lead a team planning for the initial turn on of what eventually became HP's high end SMP SuperDome systems.

Although the new 2U server was designed to be a general purpose system, the nacent Internet Service Provider market was also a target. In designing a dense system, there were tradeoffs made to accomodate the shrinking form factor. I was intrigued by what might be accomplished if we left the general purpose server requirements behind, concentrated solely on the emerging ISP environment and marketplace, and their needs for networked computing. I tried to imagine the minimal system to bring packets in a narrow pipe, run them through an application and spit them out a fat pipe. I also wondered if the high end SMP mesh backplane technologies I had been exposed to might be migrated down cost effectively.

On the flight out I had a chance to sketch up some ideas I had been mulling over for several months, and came up with a system packaging concept similar to that we would today call a blade server. My visit to the ISP data center, especially contrasting it against my experiences in traditional Enterprise data centers, reinforced my belief that this alternative packaging concept held promise.



This concept was refined over the next several months, perhaps the most interesting of all was the integration of 'payload' LAN switching directly into the product instead of remaining external to the blades. At this point, I was interested in a couple of factors relating to this concepts. First, it solved one of the problems we ran into during the development of 2U (and eventually 1U) systems: there are some components (fans, blowers, power supply components) which are difficult to shrink without encouring a stiff miniaturization penalty.

Second, there are some features (external package skins) which cost about the same regardless of their size, and increasing the number of systems in a rack only increased the number of times that customers would have to pay these fixed per systems costs. The bladed systems addressed both these challenges by obtaining the resource at an sweet spot in their pricing curves and then amortizing these costs over multiple systems.

Finally, the bladed systems offered significant improvements in time to install a server, and, perhaps more importantly, the mean time to repair. Anyone who has opened the back of a rack of 42 servers and contemplated unbundling the horse's neck of capbling to service one server can imagine how much nicer it would be to tell a data center technician "take this blade, go to the data center, when you find a blade with a blinking yellow light pull it out and put this in instead."

Eight months later, in mid-1999 I was back in Fairfax county, on a customer visit swing through the emerging category of xSP (Internet, Application, Hosting) service providers.

At the time it was a hotbed of activity during the initial expansion of the Internet economy. Almost every xSP data center manager I met with had changed jobs in the last 12 months, riding the bubble around the beltway in an upward spiral of salaries, bonuses, and stock options. Competition among these providers for customers was no less fierce.

Cluster 2002

Dominant xSP Concerns – Circa 1999

Origins

Time-to-Deployment - Competition is cut-throat. How quickly can a new service be deployed in the physical plant?

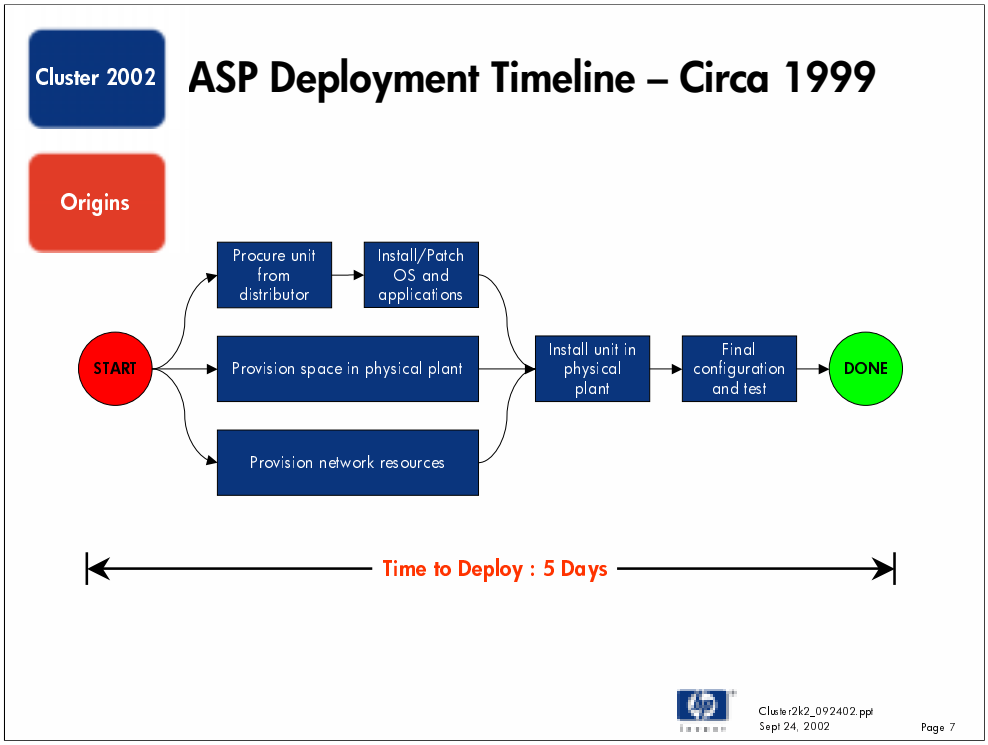
System Density - The data center incurs costs in $/area and recovers costs in systems/area. How many systems can be burdened on each square foot of data center?

Moving Customers up the price list - Currently shared hosting only sells at the introductory service level. Moving customers up the price list requires dedicated hardware

Cluster2k2_092402.ppt
Sept 24, 2002
Page 6

There were three predominant themes common to all my visits at that time, all of which had a significant influence on the first generation of blade systems.

First, the cut throat competition meant that a new metric, time-to-deployment, was key to driving growth in the customer base.



For example, at that time, a cutting edge ASP was able to deliver a new hosted application server in five days and were looking to shave off half hours from that time.

Cluster 2002

Origins

Dominant xSP Concerns – Circa 1999

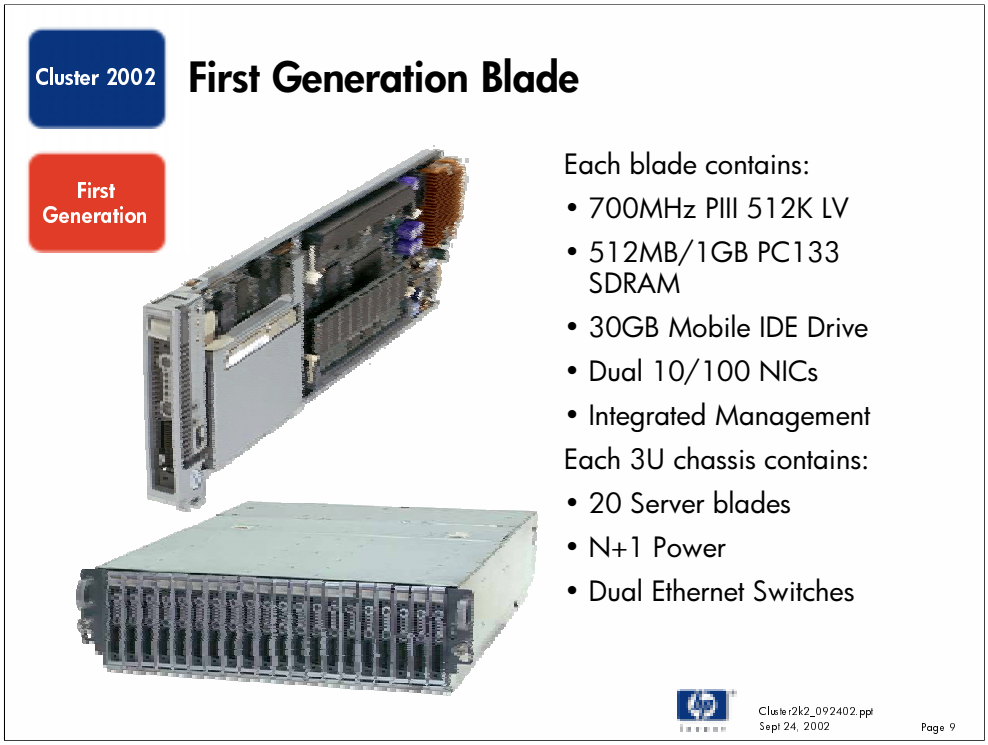
Time-to-Deployment - Competition is cut-throat. How quickly can a new service be deployed in the physical plant?

System Density - The data center incurs costs in $/area and recovers costs in systems/area. How many systems can be burdened on each square foot of data center?

Moving Customers up the price list - Currently shared hosting only sells at the introductory service level. Moving customers up the price list requires dedicated hardware

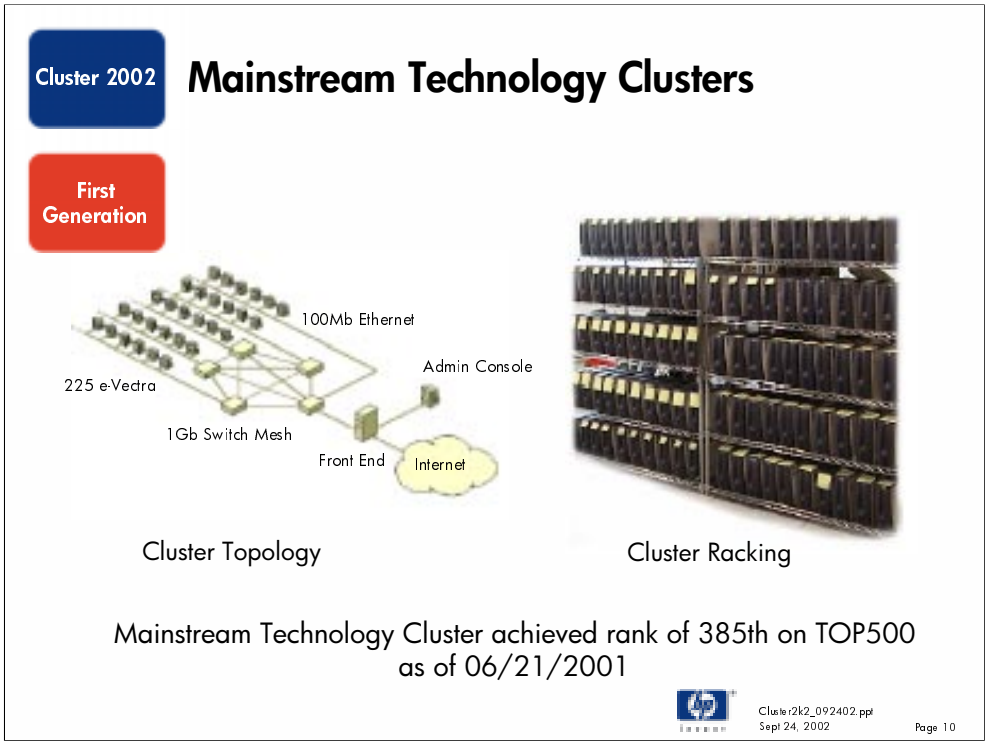
Cluster2k2_092402.ppt
Sept 24, 2002
Page 8

The second was that the xSP data center incurred costs per square foot of managed data center and recovered revenue in systems per square foot. The final common theme was that virtual hosting on a shared system was already a lost leader, and real revenue required dedicated hardware.



As the internet economy continued to expand, these forces combined to shape the first of the bladed systems introductions. In order to maximize the number of systems, these blades were characterized by reduced power processors, limited memory capacity, lap-top style disks and 100Mb networking. But, this was well matched to the generally static web traffic that these servers were tasked with. And at this point in the bubble, their was enough venture capital and 401k money coming into both the bladed system startups, the xSPs and the dot.coms that progress to design blades optimized for the network edge application continued.

This example of an HP Proliant BL e-class blade is indicative of the types of technologies that can be expected in a first generation blade available from several vendors: processors re-used from mobile systems or specifically adapted to trade off performance against power dissipation. Limited memory array. Mobile style embedded drives. Dual 10/100 networking. No high performance storage attach.



As a quick aside, It should be noted that at even this level of performance, bladed systems could be a substantial asset to the creation of cost effective, easily maintainable clusters. In October 2000, Richard Bruno, a research at HP Labs in Grenoble, who will be presenting later today, teamed with INRIA Rhone-Alpes to contruct a cluster using 225 HP e-Vectra business desktop PCs and HP Procurve ethernet switches and reached the 385th rank in he TOP500 list by June 2001. The individual e-Vectra memory, CPU, and network configuration was very similar to the first generation of blades. But, imagine how much more attractive a single rack of blade servers would be to a data center manager than than the open frame rack and hand wiring invovled in the setup of a 'mainstream technology' cluster.

Cluster 2002

Rapid Deployment Tools

First Generation

```
[root@imageserver]# getimage -g my-golden-client -image web_server_image_v1
This program will get the "web_server_image_v1" system image
from "my-golden-client"
making the assumption that all filesystems considered part
of the system image are using ext2, ext3, or reiserfs.
This program will not get /proc, NFS, or other filesystems
not mentioned above.
See "getimage -help" for command line options.
Continue? ([y]/n): y
Retrieving /etc/systemimager/mounted_filesystems from mygolden-
client to check for mounted filesystems...
------------- my-goldenclient
mounted_filesystems RETRIEVAL PROGRESS -------------
receiving file list ... done
/var/spool/systemimager/images/web_server_image_v1/etc/systemimager/mounted_wrot
e 132 bytes read 294 bytes 852.00 bytes/sec
total size is 180 speedup is 0.42
------------- my-goldenclient
mounted_filesystems RETRIEVAL FINISHED -------------
Retrieving image web_server_image_v1 from my-golden-client
------------- web_server_image_v1 IMAGE RETRIEVAL PROGRESS --
-----------
receiving file list ... done
```

Cluster2k2_092402.ppt
Sept 24, 2002
Page 11

Offering maximum systems/square foot of dedicated hardware, the blade systems seemed poised to maximize the revenue return per square foot of a network edge xSP data center. And, in order to minimize the time-to-deployment, most blade system provides were also beginning to offer provisioning software to allow for simplified mastering of the on-blade OS images via a central management station. For low end, low bandwidth applications, blades now had traction in every part of the time-to-deployment problem that was first evident in 1999.

However, there were some inhibitors. At the time, single chip 10/100 ethernet switching technology was only just becoming commodity. As a result, some first generation designs lack an integrated switch. For these systems the promise of easy deployment and maintenance of systems was true at the blade level, but not at the chassis level.

Second, the storage subsystems included in the first generation blades based on mobile style drives were more expensive, lower performance, and lower reliability than their general purpose server counterparts. The prevalent 10/100 networking didn't really have the bandwidth to allow for network attached storage, and no first generation blades had storage area network interfaces. Finally, the blades were limited in overall CPU performance and memory array size due to power and space considerations.

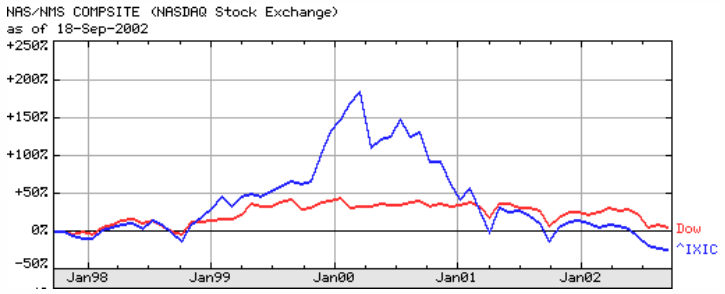
Another problem was that some designs provided little more for manageability than a command line or GUI front end to manage the images for LAN boot protocols. For those designs without a solid, blade-centric manageability model, the nearly 6X increase in density only exacerbated Server Sprawl, the proliferation and deployment of massive numbers of systems that can lead to acute manageability problems and eventually to unsustainable deployments.

Still, as long as web sites need to be provisioned as quickly as before ...

Cluster 2002

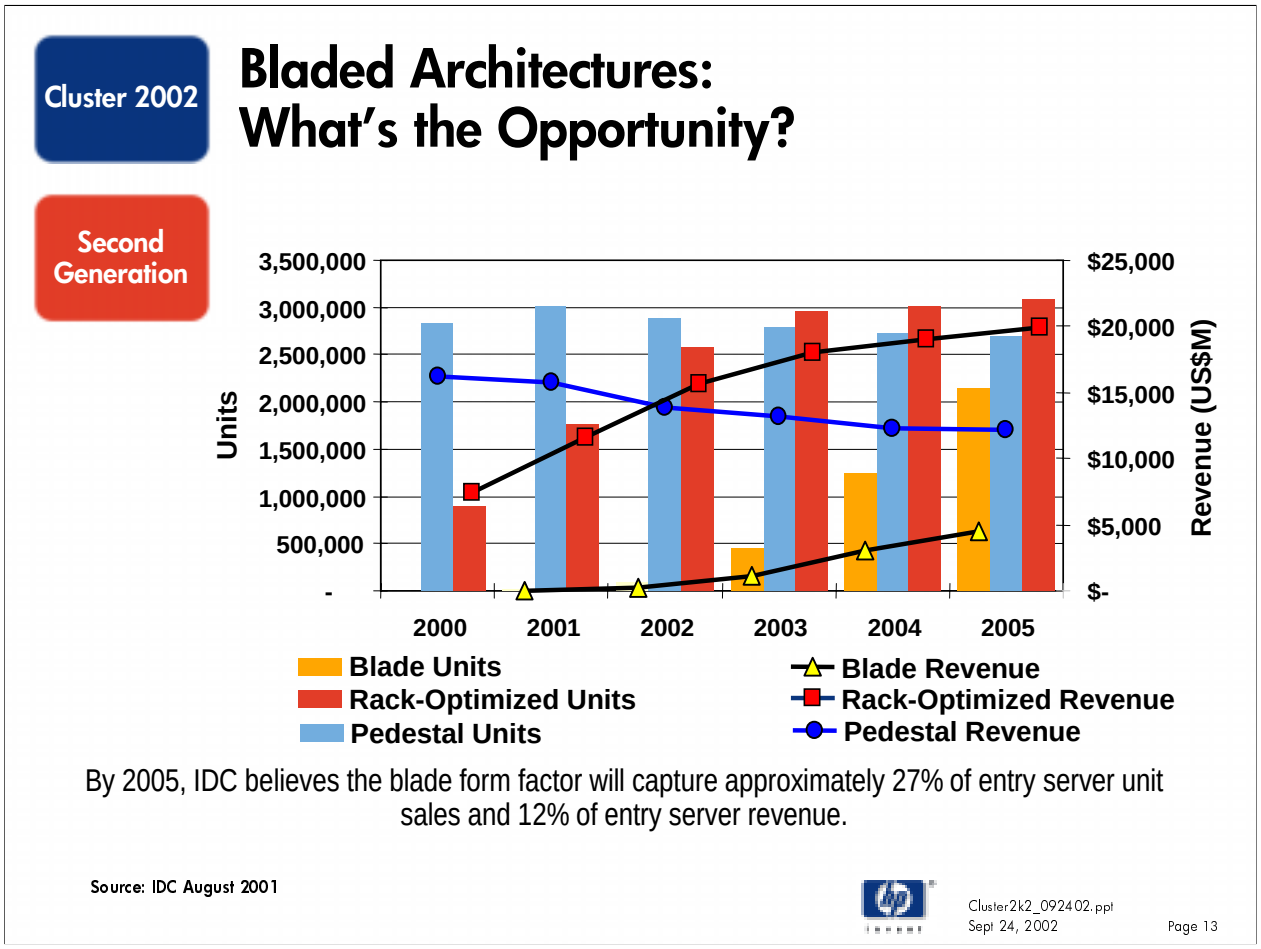
Limitations

What New Economy?



Pop! Did we say new economy? With the very rapid decay of the internet bubble economy, the demand for rapid build out of thin, dense servers was quickly eroded. Along with failing dot.coms and xSPs, many of the startups designing bladed systems also found themselves without a market.

But, this shakeout did not dissuade the major vendors from continuing to explore the density and total cost of ownership advantages of blades.
In fact, although thin blades are a great way to deploy network edge servers, the cost of ownership advantages of blades can be a benefit in high performance systems as well.

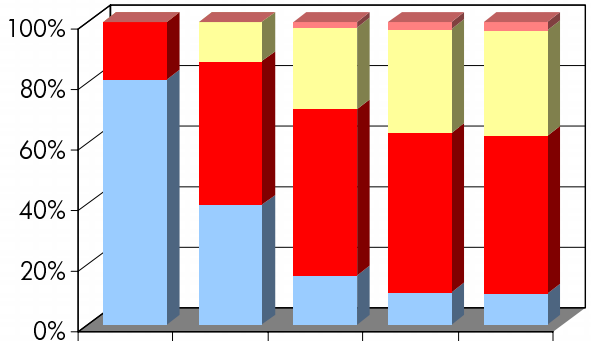


As is shown in this IDC market projection, the penetration of blade into the server marketplace is anticipated to ramp steadily over the next several years, culminating in a 27% penetration rate in 2005.

Cluster 2002

Second Generation

Bladed Architectures: Where Are the Sweet Spots?



	2001	2002	2003	2004	2005
8-way	0%	0%	2%	2%	3%
4-way	0%	13%	27%	35%	35%
2-way	19%	47%	55%	52%	52%
1-way	81%	39%	16%	11%	10%

2005
$4.5 billion

Share of bladed server revenue quickly migrates to the 2 & 4 processor "flavors" as volume ramps and commoditization occurs

Source: IDC August 2001

Cluster2k2_092402.ppt
Sept 24, 2002
Page 14

But, perhaps also more interesting is this IDC product segmentation projection which shows the penetration of blades with dual, quad, and even octal SMP configurations in the same timeframe. Clearly these will not be the mobile technology enabled blades of the first generation, but what will they look like?

Cluster 2002

Second Generation Blade

Second Generation

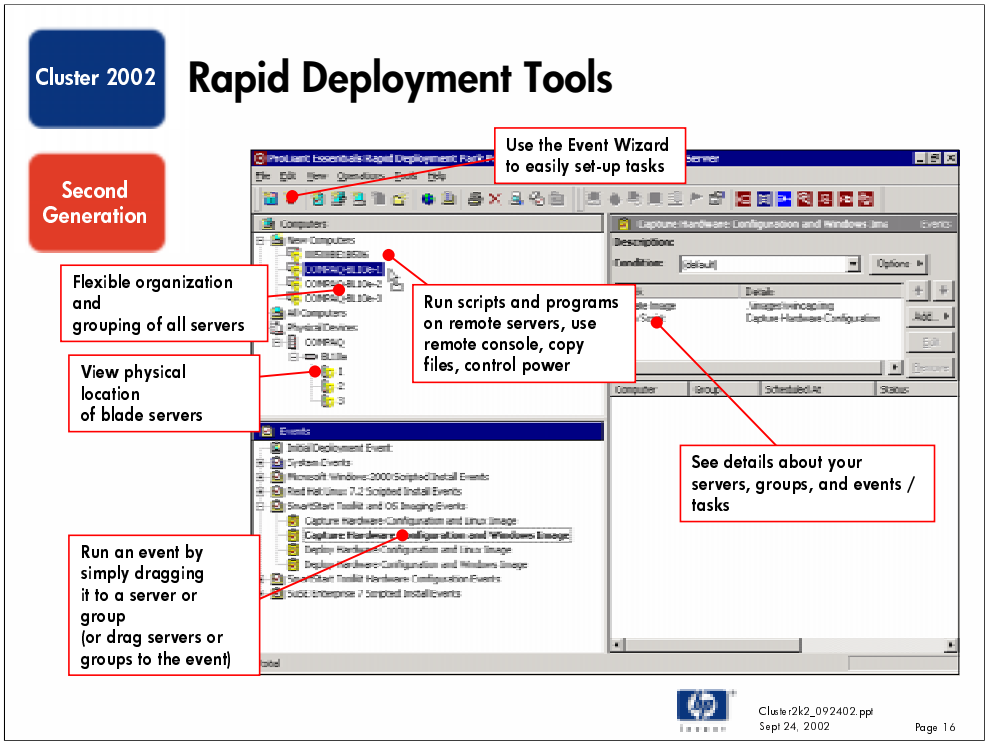
- Processors: Mobile/LV/ULV → Mainline server Processors
- SMP: Single → Dual (or more) Processors
- Memory sizes: Sub-1GB per processors → Greater than 1GB per processor
- Ethernet NICs: Dual 10/100 → Multiple 10/100/1000
- Manageability: rack-optimized server leveraged → blade-aware
- Local Storage: low performance mobile → high performance, high reliability server storage
- Remote Storage: Low performance NAS → high performance NAS or SAN
- Chassis form factor: 3U → 6U
- Blades form factor increase in depth, width, and height

Cluster2k2_092402.ppt
Sept 24, 2002 Page 15

This example of an HP Proliant BL p-class blade is generally indicative of what most major vendors are currently delivering or planning for this next generation of high performance blades. The general theme of these designs is that the same technology present in high performance 1U/2U rack optimized servers will be available in this form factor. In particular:

- Mobile/LV/ULV processors migrate to mainline server processors
- Single processors migrate to Dual (or more) Processors
- Memory sizes migrate from Sub-1GB per processors to greater than 1GB per processor
- Ethernet NICs migrate from Dual 10/100 to Multiple 10/100/1000
- Manageability migrates from rack-optimized server leveraged to blade-specific
- Local Storage migrates from low performance mobile to high performance, high reliability server storage
- Remote Storage is enabled via high performance NAS or SAN
- Chassis form factor increases from 3U to 6U
- Blades form factor increase in depth, width, and height

This provides a snapshot of currently available high performance bladed systems. It is interesting that although the first generation system designs responded to the economic drivers of the internet bubble, these mainstream performance systems, through their vastly decreased MTTR, their excellent blade-centric management integration, and their amortization of shared resources over multiple systems, will drive blade adoption past the network edge and into the enterprise data center. As the bladed systems model matures, and especially as innovation continues in the management software infrastructure, modular hardware will continue to penetrate traditional rack optimized server market. I say modular hardware, since the label blade will become an increasingly inaccurate metaphor for future higher performance modules

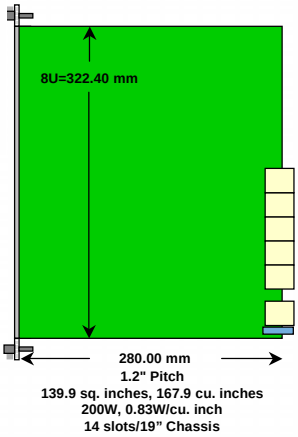


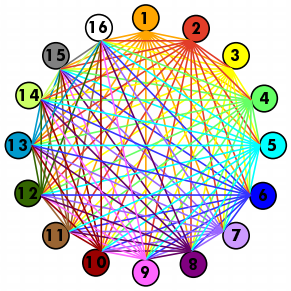
Along with Second generation blades will come second generation rapid deployment tools. This screenshot of HP's Proliant Essential Rapid Deployment Pack shows the advances in blade-specific management tools that can prevent Server Sprawl.

Cluster 2002

Futures

Advanced Telecom Computing Architecture





Each blade contains:

- A 200W Thermal Envelope (enables Dual slot / DP IPF or QP IA-32)
- Fully connected Mesh Data Transport Protocol Agnostic Fabric. Each Link can support, for example:
  - 1 4X Infiniband link/4 1X Infiniband links
  - 4 1000BX links
  - 1 10GbE XAUI link
  - 4 2Gpbs FC links
- Dedicated Management Dual Star 10/100/1000 Ethernet
- 10 pair Adjacent Slot Update bus
- Redundant blade-centric hardware management bus
- N+1 Redundant Power
- Cost focused module mechanical design

Cluster2k2_092402.ppt
Sept 24, 2002 Page 17

So blades systems are now capable of bringing their total cost of ownership and rapid deployment characteristics to the dense, horizontally scaled network edge as well as the vertically scaled data center. Where might they go next?

A possible leading indicator of trends in modular hardware features is an effort that I have been driving for HP in the last year, the Advanced Telecommunication Computing Architecture or ATCA draft standard. ATCA is a draft standard of the PCI Industrial Computer Manufacturers Group targeted at bringing next generation modular design standards to a converged communications compute market and is slated for adoption by year end. This is a summary of some features of the current draft definition of ATCA. While some aspects of the draft standard are particularly optimized to the converged communications market, many of the features of the standard could be leveraged into a third generation blade.

Cluster 2002

Futures

ATCA Target Serial I/O Standard Comparisons

Serial Standard	Data rate per Channel (Gbps)	Baud rate per Channel (Gbps)	Ref Clock (MHz)	Pairs/Channel Channel Configs
Infiniband	2.0	2.5	125	2 1X, 2X, 4X, 12X
1Gb Ethernet 1000BASE-CX	1.0	1.25	62.5	2 1X
10Gb Ethernet XAUI	2.5	3.125	156.25	2 4X
Fibre Channel	0.85 / 1.7	1.06 / 2.12	53.125 / 106.25	2 1X
Serial ATA	1.2	1.5	75	2 1X
Serial RapidIO	2.5	3.125	156.25	2 1X,4X
PCI Express	2.0	2.5	125	2 1X-32X

Cluster2k2_092402.ppt
Sept 24, 2002
Page 18

One interesting aspect of ATCA is that while it is optimized around modular hardware interconnected via a fully connected mesh of high speed serial I/O connects, it doesn't specify which I/O connects are utilized. The integrated managements systems handles the possible heterogeneous mixture of technologies. This chart shows a current snapshot of ATCA compatible I/O connects. All these I/O connects have very similar LVDS signaling, similar bit counts and bit rates, and differ primarily in the lower layer protocol enabling and end use models.

Now this brings up an interesting question: what will the fabric of choice be for the next generation bladed system. Ethernet was the obvious choice for first generation systems, since their market called for and they only had the capacity for network connectivity. The second generation is still predominately ethernet based, although now with the capacity for high performance storage interconnect, fiber channel has been discussed. Also, from Infiniband silicon vendors have demonstrated bladed systems.

At this point, it is difficult to deny the economy of scale being afforded to ethernet is a distinct advantage, but other I/O connects do have feature advantages and efficiencies that could be useful in bladed systems. There is also the distinct possibility that the partial physical layer convergence will continue and could yield physical unification in the 10Gb (XAUI) or 40Gb (XAUI DDR) generations.

This is quite an advancement rate. From pencil sketches in late 1998 to draft standards for high performance modules with full power thermal envelopes fully connected by meshes with multi-terabit class bandwidth built from mainstream commodity processors and I/O technologies in late 2002.

On behalf of Hewlett-Packard, I'd like to thank the IEEE Computer Society and the Task Force on Cluster Computing for the opportunity to present to you today.