Computing Servers

(V1.2)

Arquitecturas Clúster

ETSI Informática

Depto. de Arquitectura de Computadores
Universidad de Málaga

Sections

- The Race for High Density
- Server Form Factors

Contents

THE RACE FOR HIGH DENSITY

The Race for High Density

- The problem of high density:
  - The demand for new services (apps) is continuous and the number of servers required increases continuously.
  - The area of the **cold room** in the data center is limited. Enlarging the room is very expensive in case there is available room.

- The objective:

  To place as much CPU sockets, RAM slots and storage space as possible in each square meter of the cold room.

- The challenges are:
  - Equipment size reduction
  - Electric power density, delivery (power cabling) and available budget
  - Heat density limits and limited cooling capacity
  - Network bandwidth and delivery (network cabling)

Equipment Size Reduction

- Increase in the capacity and reduction in the form factor of every hardware component
  - CPU
    - More CPU sockets in motherboard
    - More cores per CPU socket
    - L1/L2 caches were integrated inside CPU die years ago; on-board chipset is following the same path.
  - Memory
    - Larger DIMM modules
    - More DIMM sockets in motherboard
  - Storage
    - More information density for magnetic disks and tapes
    - 2.5” Small Form Factor HDDs (SFF) and mPCIe form factor SDDs replace 3.5” Large Form Factor HDDs (LFF)
    - More disks per rack unit in servers and high density vaults

CPU cores density

The Race for High Density

4 CPU x 12
(48 cores)

4 CPU x 16
(64 cores)

RAM Density

The Race for High Density

16 DIMM slots in half-width (max 512GB RAM)

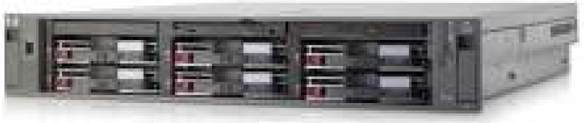
48 DIMM slots in 2U (max 3TB RAM)

96 DIMM slots in 4U (max 6TB RAM)

Disk density

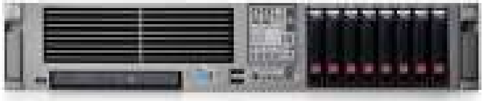
The Race for High Density

6 LFF server

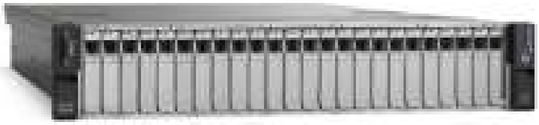


16 LFF vault

48 SFF vault



8 SFF server



24 SFFvault

60 SFF vault

SERVER FORM FACTORS

Servidores Tipo Torre

- Pros
  - Entry-level price is lower
  - No rack is needed
- Cons
  - Bad room scalability
  - Ugly cabling
  - Low-end servers without RAS features
  - Low CPU core and RAM density.

Server Form Factors

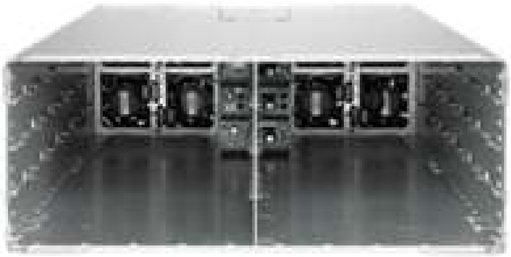
Full Width Servers (Pizza boxes)

Server Form Factors

- Pros
  - More density (up to 42 servers per rack).
  - Smart cabling running along rack backside.
  - High end servers need more room for CPU sockets, DIMM sockets, disk bays and PCIe card slots (up to 4U, 8U or even 10U).
- Cons
  - The rack has an high entry-level price even for one server.
  - Rack mounted servers are middle to high end machines with RAS features. Server entry-level price is also high.
- Storage expansion with rack mounted disks vaults
  - 1U servers only require 2 system disks in RAID1
  - Data disks are in separate rack modules
  - Disk vault can implement RAID or more advanced storage features (deduplication, snapshots, compression, encryption, etc)
  - Storage vaults can be attached in several ways:
    - DAS to a single server (Direct Attached Storage)
    - SAN to multiple servers (Storage Area Network)
    - NAS to multiple servers (Network Attached Storage)

Half Width

- Two servers side to side in 1U
- Power Supply Units (PSU) and fans are shared with other servers to save space
- Pros
  - More density (up to 84 servers in a rack)
  - Modular: 1U, 2U, 4U servers available with more room for CPU sockets, DIMM sockets, disk bays and PCIe card slots (up to 4U, 8U or even 10U)
  - Centralized management of all servers through chassis
- Cons
  - A chassis is required to hold servers, PSUs and fans with a central wall to divide width in two halves
  - High density formats tend to be more expensive for the same resources
  - Components are not compatible with other manufacturers



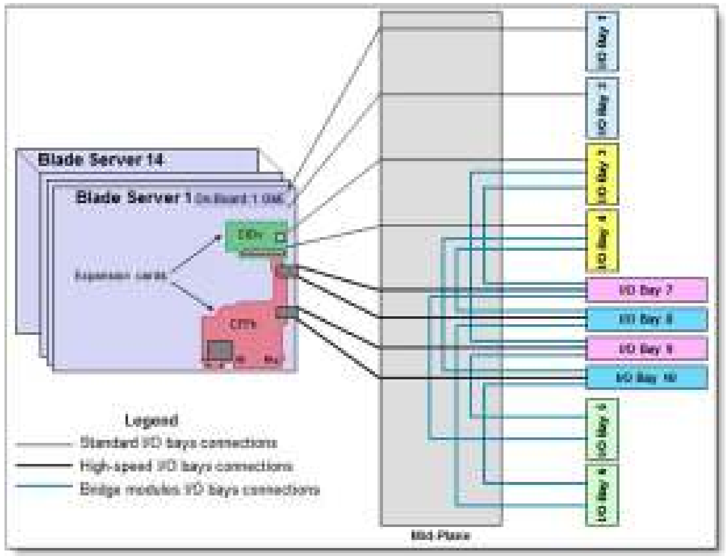
Blade Centers

Server Form Factors

- Even more density than half width servers
- All server external I/O connectors are missing to reduce server size
- Chassis contains:
  - PSUs and fans
  - I/O switches (mezzanine) with external I/O connectors (Ethernet, Infiniband, Fibre Channel)
  - Centralized management controller
  - Unfrequently used devices: USB ports, VGA, DVD, power and reset button, status lights, etc.
- Pros
  - Even higher density (14 to 16 servers in 6U).
  - Less cabling (chassis implements internal networking between servers)
  - Better resource scalability
  - Blade servers should be cheaper as they lack many shared components
- Cons
  - Higher entry-level price for the chassis
  - Components are not compatible with other manufacturers
  - Server I/O adapters and internal switches are specific for a blade center model and thus more expensive

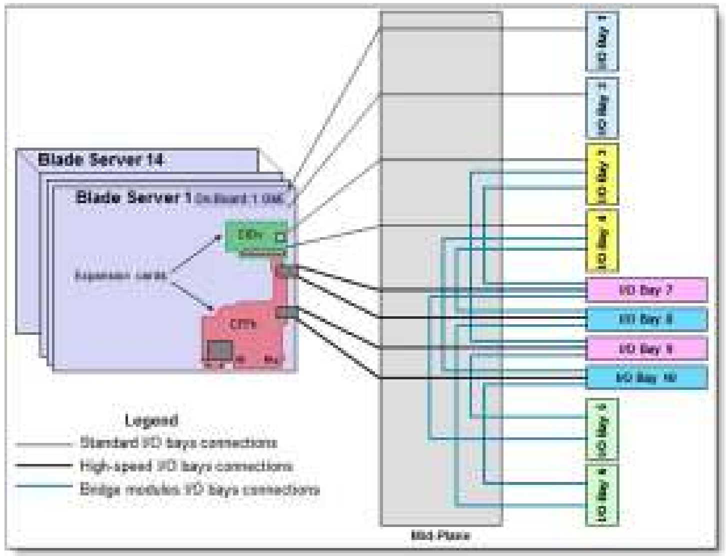
BladeCenter I/O Architecture

- Blade I/O architecture is responsible of connecting I/O interfaces in servers with external devices
  - Blade servers need I/O **adapter modules** with **internal connectors** to mid-plane
  - Chassis bays hold p**ass-through or bridge modules** with **external connectors**
  - Chassis mid-plane routes signals from server adapters to bridge or pass-through modules in bays
  - External device cables attach to external connectors of modules in bays



Example:
IBM BladeCenter H Chassis

- BladeCenter Bays
  - *Bays 1 and 2 support only standard Ethernet-compatible I/O modules. These bays are routed internally to the onboard Ethernet controllers on the blades.*
  - *Bays 3 and 4 can be used either for standard switch or pass-through modules (such as 8 Gb Fibre Channel or Gigabit Ethernet modules) or for bridge modules. These bays are routed internally to the CIOv connector on the blades.*
  - *Bays 5 and 6 are dedicated for bridge modules only and do not directly connect to the blade bays. Bridge modules provide links to the I/O bays 7 - 10 and can be used as additional outputs for I/O modules in those bays. If I/O bays 3 and 4 are used for bridge modules, they are not directly connected to the blades, and bay 3 provides redundancy for bay 5, and bay 4 provides redundancy for bay 6.*
  - *I/O bays 7 - 10 are used for high-speed switch modules such as the IBM Virtual Fabric 10 Gb Switch Module or Cisco Nexus 4001I Switch Module. These bays are routed internally through midplane connectors to the ports on CFFh expansion cards (with HS23 blade, I/O bays 7 and 9 are routed to the integrated 10GbE ports on HS23 through LOM Interposer Card). I/O bays 7 - 10 can also be converted to the standard I/O bays with the Multi-Switch Interconnect Module (MSIM).*



→ IBM BladeCenter H I/O Architecture

→ IBM BladeCenter Server HS23 I/O Expansion Options

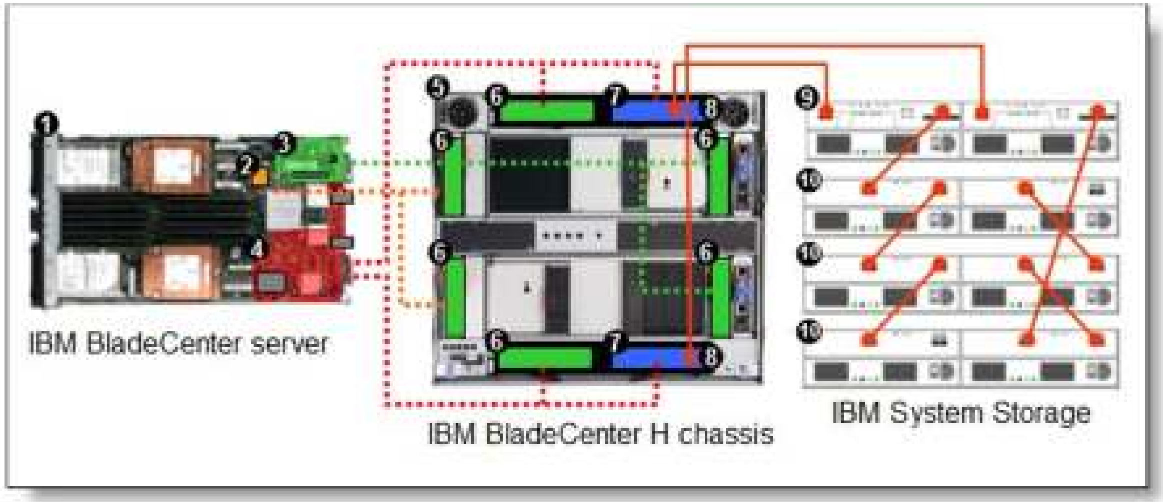
→ Emulex 8Gb Fibre Channel Expansion Card (CIOv) for IBM BladeCenter

→ QLogic Ethernet and 8 Gb Fibre Channel Expansion Card (CFFh) for IBM BladeCenter

→ QLogic 20-Port 8 Gb and 4/8 Gb SAN Switch Modules for IBM BladeCenter

Example:
IBM BladeCenter H Chassis

Server Form Factors



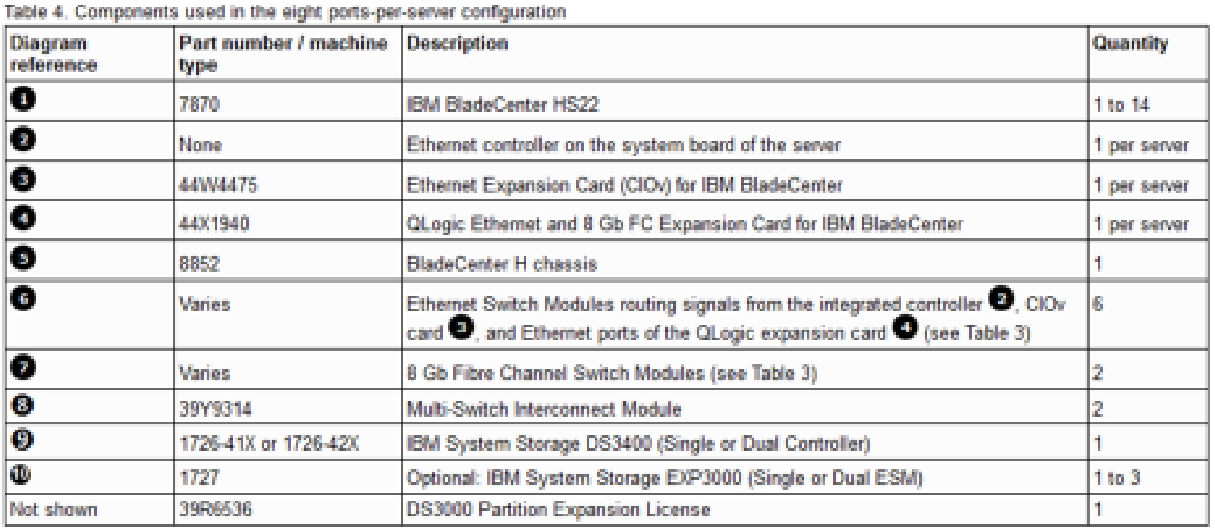


Table 4. Components used in the eight ports-per-server configuration

Diagram reference	Part number / machine type	Description	Quantity
❶	7870	IBM BladeCenter HS22	1 to 14
❷	None	Ethernet controller on the system board of the server	1 per server
❸	44W4475	Ethernet Expansion Card (CIOv) for IBM BladeCenter	1 per server
❹	44X1940	QLogic Ethernet and 8 Gb FC Expansion Card for IBM BladeCenter	1 per server
❺	8852	BladeCenter H chassis	1
❻	Varies	Ethernet Switch Modules routing signals from the integrated controller ❷, CIOv card ❸, and Ethernet ports of the QLogic expansion card ❹ (see Table 3)	6
❼	Varies	8 Gb Fibre Channel Switch Modules (see Table 3)	2
❽	39Y9314	Multi-Switch Interconnect Module	2
❾	1726-41X or 1726-42X	IBM System Storage DS3400 (Single or Dual Controller)	1
❿	1727	Optional: IBM System Storage EXP3000 (Single or Dual ESM)	1 to 3
Not shown	39R6536	DS3000 Partition Expansion License	1

Mainframes

- The traditional concept of a very large shared memory computer
  - Usually a NUMA architecture
    - Several CPU modules with DDR RAM
    - Several I/O modules with PCIe slots
    - A proprietary memory interconnection network
  - They run a **single image Operating System** (boots as a single machine where all processors see a unique shared memory space)
- They span over one ore more standard size racks to hold all CPU and I/O modules
  - Racks need to be close to each other as the interconnection network uses custom delicate cooper cables
- Examples: SGI Altix and SGI UV series, HP SuperDome, IBM System z

Server Form Factors

Mainframes

- Pros
  - Simpler to admin: single machine with a lot of resources
  - Applications scale up with very simple programming model (threads with shared memory)
    - Ex: SGI UV scales up to 2048 cores with up to 16TB of RAM.
  - Integrated Software and Hardware RAS features
- Cons
  - Far more expensive than a cluster

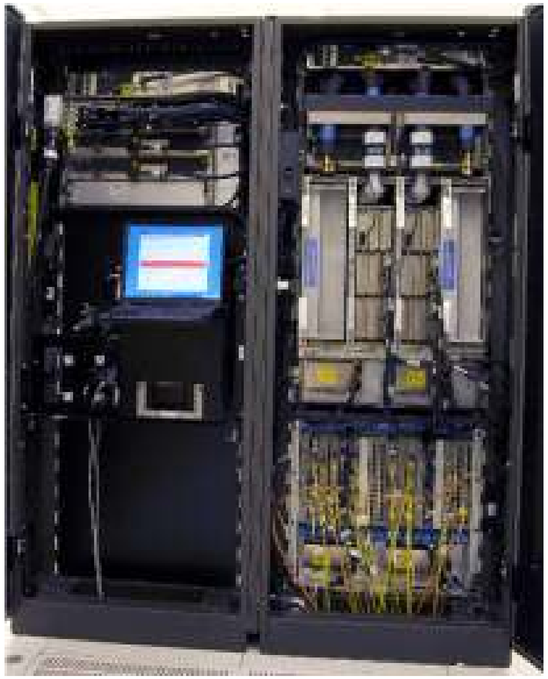
Mainframes

- Management firmware allows to partition the hardware in multiple partitions or **static virtual machines**
  - Can operate as a single large NUMA machine.
  - Can operate as a set of smaller NUMA machines.
  - Virtual machines are statically defined and a complete reboot is needed to change partitioning.

- An VMM (hipervisor) can dynamically manage multiple virtual machines:
  - Virtual machines are the base for many RAS features
    1. Faults are notified to hipervisor.
    2. Hipervisor reconfigures virtual machine hardware: fault hardware is disabled and replaced by space hardware (CPU, RAM, I/O paths).
    3. If error is not transparently recoverable, virtual machine is automatically halted, reconfigured and rebooted.

Mainframes are not dead: IBM System z

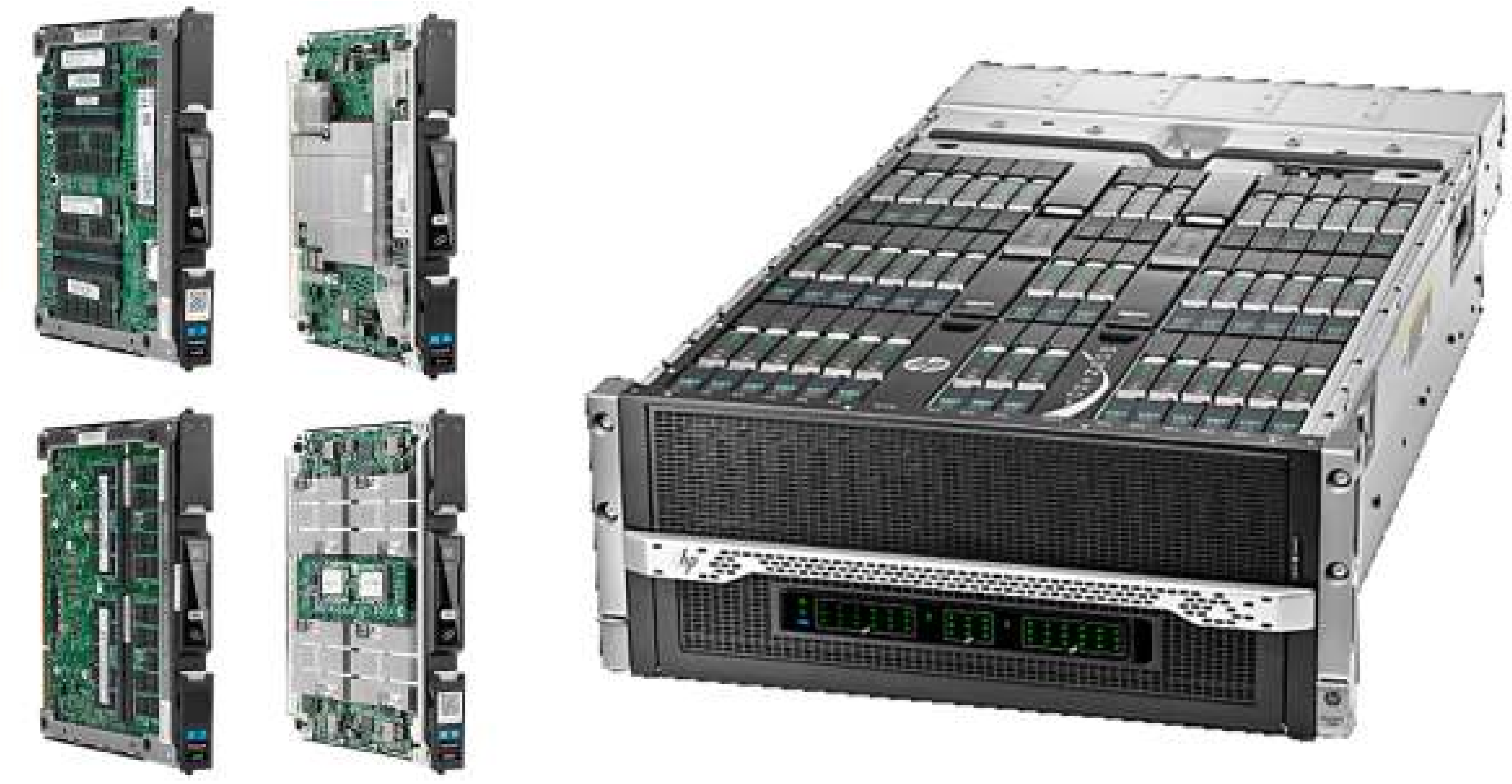
- ♦ Mainframes in perspective: A classic going strong



Server Form Factors

HP Moonshot

Server Form Factors



HP Moonshot

- 45 **independent, low power, small** **cartridges** without external storage in a 4.5U chassis
  - Cartridge provides:
    - CPU + RAM + Network (2xNIC) + local disk (1xSATA)
    - Up to 4 servers per cartridge
  - Chassis provides:
    - Power, cooling, remote management.
    - 2 Network switches with:
      - 45 ports (1GbE or 10GbE)
      - 6 SFP or 4 QSFP uplinks (several configurations: 6x10GbE, 4x40GbE, 16x10GbE)
- Target:
  - Providers of **massive hosting** for private servers (up to 180 servers per 4.5U chassis)
  - Clusters of independent systems for **distributed processing** of **massive data**.

HP Moonshot

- Specialized server cartridges:
  - ARM Based
    - **Web caching**: 1 ARMx8 cores, 64GB DDR3, up to 480GB M.2 SSD
    - **Real-Time DSP and Telco Infrastructure**: 1 ARM x 4 cores + 8xC66 DSPs, 32GB DDR3, 64GB SSD
  - Atom Based
    - **Web Infrastructure**: 1 Atom x 8 cores, 32GB DDR3, up to 1TB HDD or 240GB SDD
    - **Web hosting**: 4 x (1 Atom x 8 cores + 16GB + 1x 64GB iSSD)
  - x86 Based
    - **Hosted Desktop**: 4 x (1 AMD Opteron X2150 APU + 8GB RAM + 1x64GB iSSD)
    - **Application Delivery (Citrix) and Video Transcoding**: Xeon E3-1284, 32GB DDR3, up to 480GB M.2 SSD

SPECIFIC PURPOSE SERVERS

Manufacturer Server Series