Blade system architectures for scale-out clusters

Gene Kusekoski

Linux Blade System Solutions Engineering Mgr
Hewlett-Packard

Agenda

- UNIX SMP migration to Linux blade system clusters
- Matching blade clusters to application models
- Cluster storage models
- Virtualized Linux scale out cluster examples

(Products are examples, not recommendations)

Definitions

Scalability: Ability to expand a “system” to handle increased workloads

Scale Up: Increasing compute power of an SMP server by adding CPUs and memory to the same server (vertical scaling)

Scale Out: Increasing compute power by adding servers for workloads spread across multiple systems (horizontal scaling)

Cluster: Closely coupled group of similar systems with uniform networking and storage dedicated to a common task set

Grid: Dynamic group of diverse, distributed systems with varied network connections that contribute compute cycles to a task

Traditional UNIX SMP servers

- What's to love
  - One system to manage
  - Scales well to its designed capacity
  - Access to very large uniform memory space
  - Unified networking and storage
- Challenges
  - Proprietary H/W and S/W
  - High capital costs
  - High maintenance costs
  - Potential underutilization & forklift upgrades
  - Double costs for redundancy / availability

Linux blade cluster alternative

- What's to love
  - Low cost standard building blocks
  - Industry standard & open source S/W
  - Capital $ = SMP maintenance $
  - Low maintenance costs (blade FRU)
  - Incremental sizing, redundancy
  - Independent networking and storage
- Challenges
  - Applications must be distributable
  - No large uniform memory space
  - Networks vs crossbar interconnects
  - Potentially more pieces to manage

Why blade systems?

- Integrated modular infrastructure
  - Rack it once, configure pluggable modules
  - Add/upgrade modules to follow technology
  - Up to 87% reduction in cabling
  - Shared resources, e.g., power
  - Intelligent profiling (provision by slot)
  - Provisioning / management software options
- Ideal fit for 2P scale out sweet spot
  - CPU options, including extended memory
  - Space-efficient compute density
  - Inherent compute / storage modularization

"Blade servers are altering the environment of the data center."

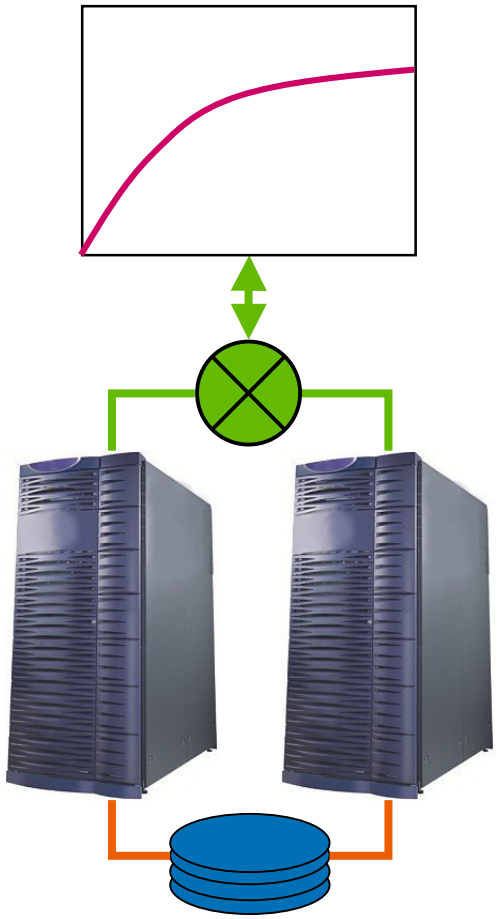
William Terrill

Senior Analyst, Burton Group

Before and after…

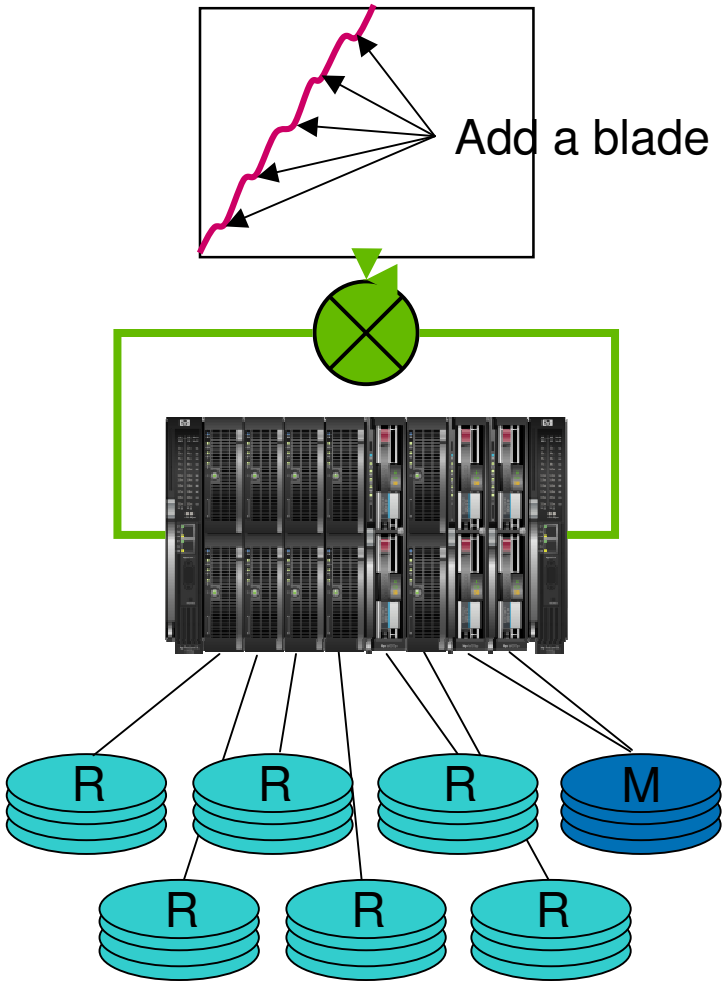
Case 1a: SMP database

- Replicated servers for failover
- Hot standby or 50% load balanced
- Writes steal from read capacity
- Storage failure takes data offline
- I/O limits benefit of SMP
- No practical way to scale



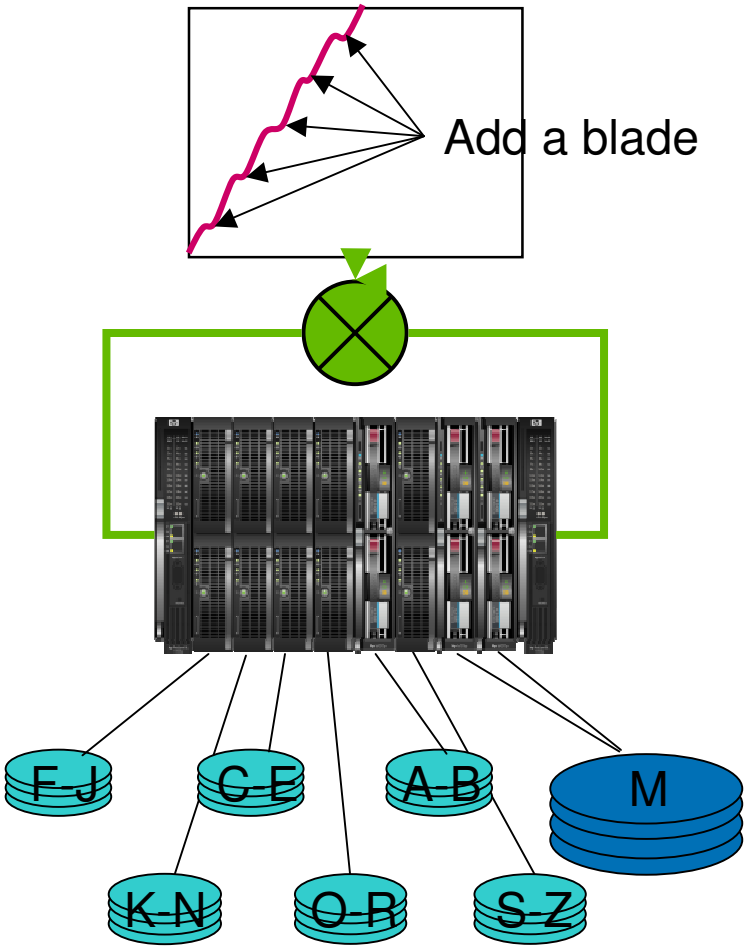
Case 1b: Replicated database cluster

- Presumes moderate size database
- Presumes heavy read access
- High availability master database
- Add blades to scale read layer
- Replicas balance CPU & I/O
- Use standard load balancing
- Service up even if main db offline
- Any failure only reduces capacity
- Principle used for Sabre service



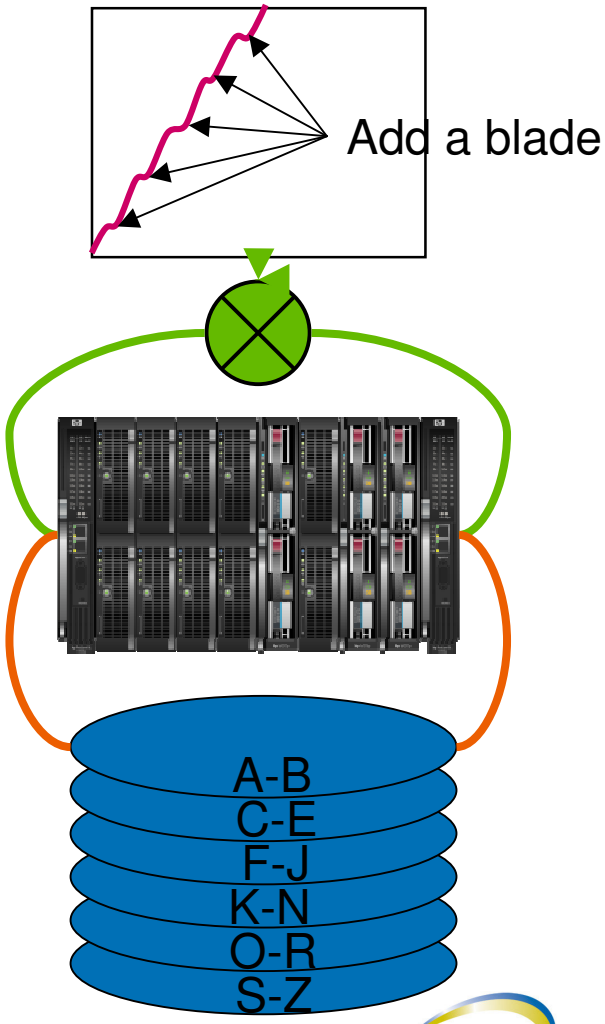
Case 1c: Federated database cluster

- Variant of replicated design
- Supports larger databases
- Availability with N+1 spare blade
- Configure storage for availability
- May be harder to load balance



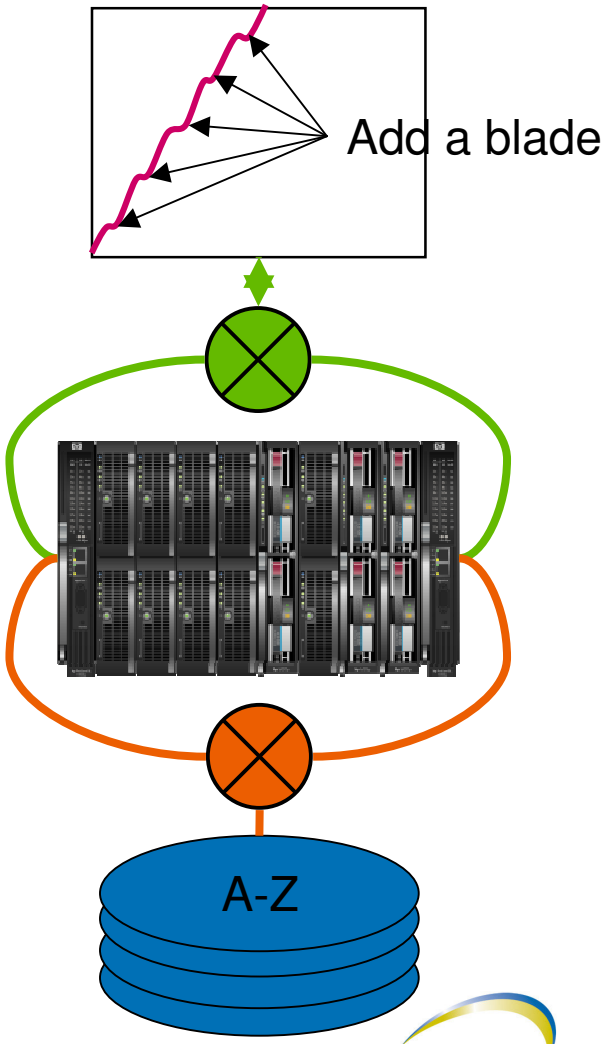
Case 1c2: Shared storage db cluster

- More flexible federated model
- Segments in shared storage LUNs
- Availability with N+1 spare blade
- Spare blade takes master or read role
- Potential for better load balancing
- Model also applies to “tier 2”
  - Web servers
  - Java app servers



Case 1d: Cluster-aware database

- Oracle 9i RAC / 10g
- Shared data on OCFS (raw)
  - Concurrent access to shared database
  - Inherent resilience to node failure
  - Expansion without db or app mods
  - OCFS for database elements only
- OS on internal DAS
- Oracle Home LUN per server
- Availability /scalability by design

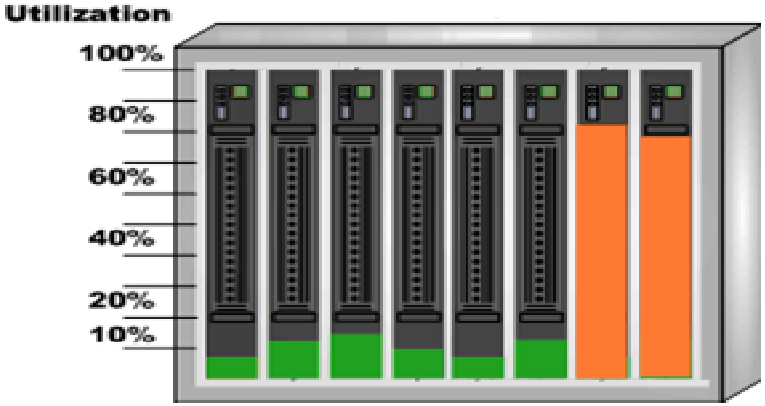


Case 2a: Multiple application SMP

- Use SMP system partitioning
  - CPU
  - Memory
  - Storage
  - Networking
- Rebalancing possible
- Resource bottlenecks possible
- Potential wasted capacity
- Forklift upgrade path
- High cost proprietary approach

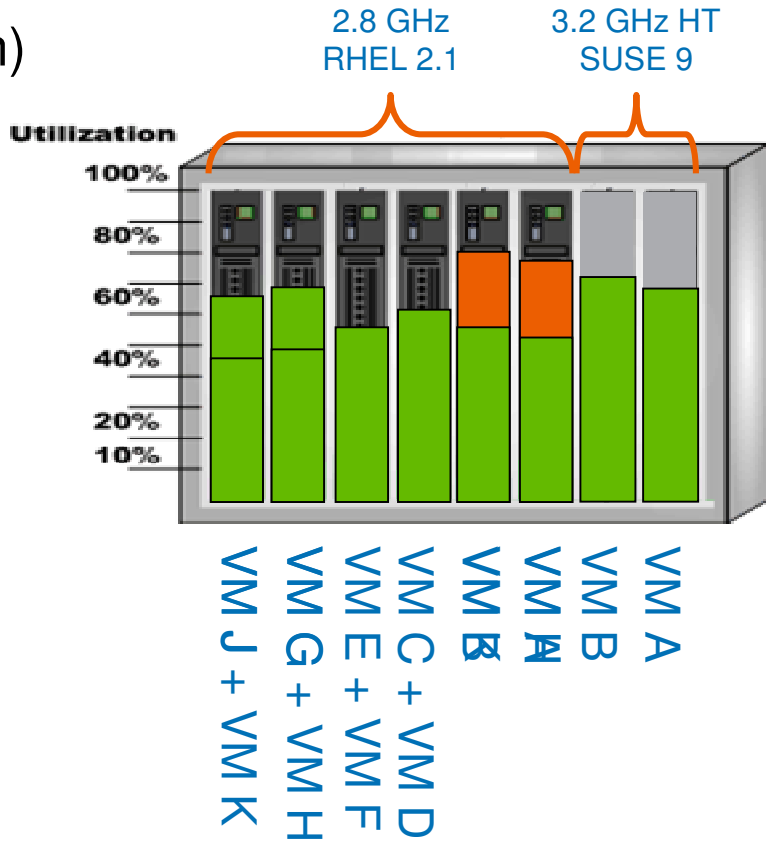
Case 2b: Application blades

- Assign applications to blades
- Low cost modular approach
- Add apps by adding blades
- Good for apps that
  - Don't play well together
  - Need different OS variants
- But…
  - May be wasteful
  - May limit some applications



Case 2c: Virtual application servers

- Virtual machine environment
  - (VMware ESX, Virtual SMP, VMotion)
- Virtual machine per application
- Add more blades as needs grow
- Manually rebalance as needed
- Migrate VMs to faster blades
- Rolling OS & app upgrades

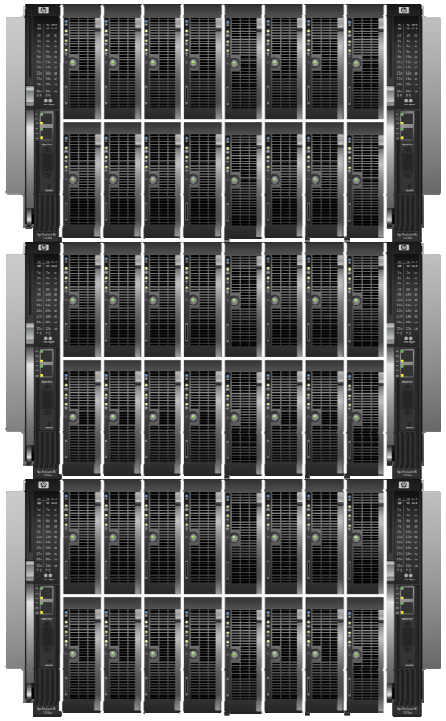


Case 3a: Heavy duty computing

- Technical & financial analysis
- Need for horsepower
- May need uniform large memory
- Modest storage & networking need
- Potential wasted capacity
- Even 128 CPUs may not be enough
- SMP scaling may be non-linear

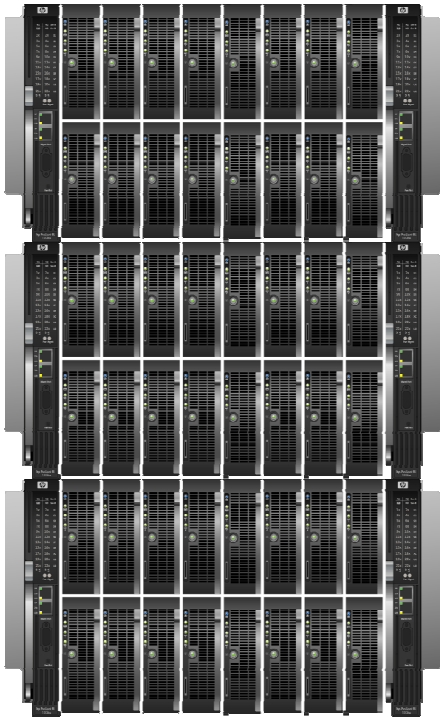
Case 3b: Simple compute cluster

- Beowulf clusters 10 years old
  - Fraction of the cost of SMP systems
  - Control + worker node design
  - Provision workers with OS only (e.g., RDP)
  - Compute power evolves with technology
- Wealth of Linux, open-source work
  - Scyld, OSCAR, Rocks, Scali kits/wizards
  - Standard interfaces (MPI, PVM)
  - Out of band mgmt hooks (e.g., iLO, IPMI)
- Ideal for high-density compute blades
- Requires parallel processing apps
  - Moderate scale out (Amdahl's Law)
  - Best on "coarse-grained" apps (rendering)



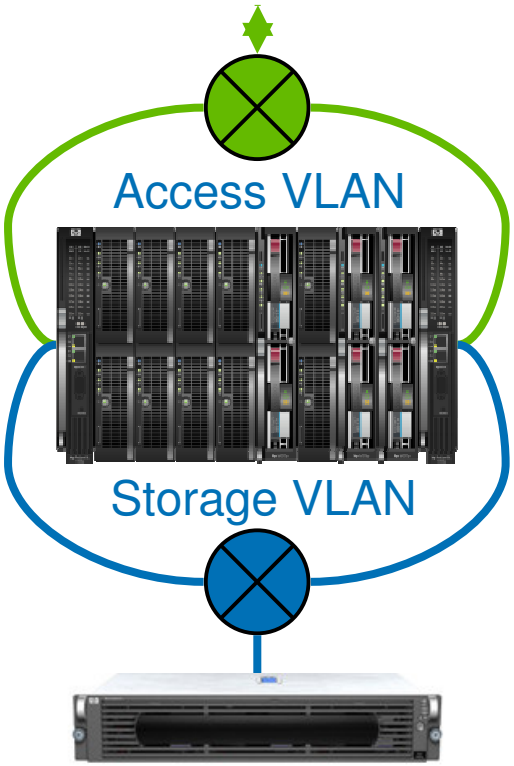
Case 3c: Virtualized compute cluster

- Accelerated by 21st century technology
  - Increasing CPU computes/watt/in^3
  - Improving interconnect speed/latency
  - Linux & open source adoption
- Single System Image (SSI) model
  - Linear scaling design / shared root
  - OpenSSI (Bruce Walker)
  - openMosix, Qlusters (Moshe Bar)
- Cluster optimizes process distribution
  - Dynamic resource monitoring & allocation
  - Boot in new capacity (no preloaded S/W)
  - Applications work as-is (inter-blade processes)
  - Parallel processing applications work better



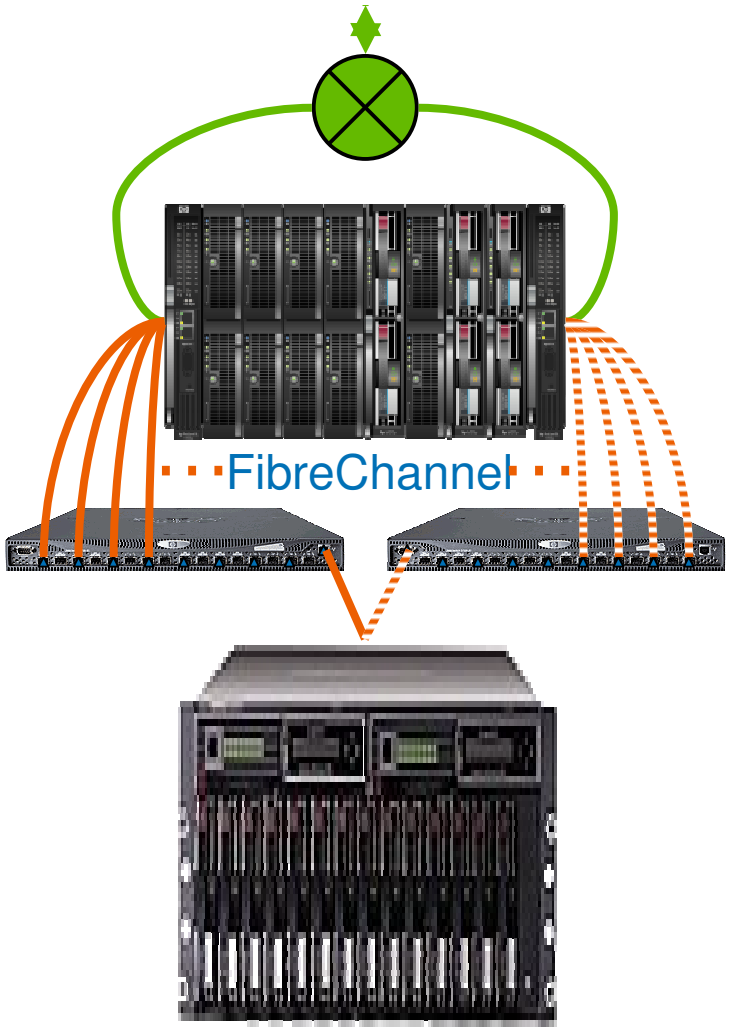
Storage models: NAS/NFS

- Lowest cost shared storage
- Standard network platform
- Long distance interconnects
- Replicate anywhere on network
- NAS appliance or NFS service
- File based with lock management
- Moderate performance
- Good for compute clusters



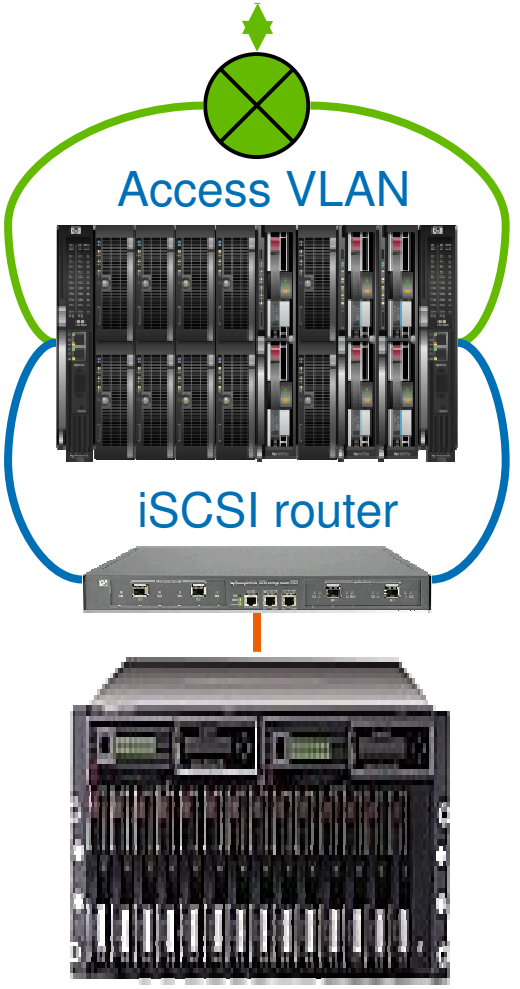
Storage models: SAN

- Premium shared storage
- Specialized interconnect
  - FC adapter per blade
  - Short distance interconnects
  - Redundancy options
  - Long distance bridge possible
- Basic to high end models
- Block based
  - LUNs replace DAS
  - No file lock management
- High performance
- Ideal for database applications



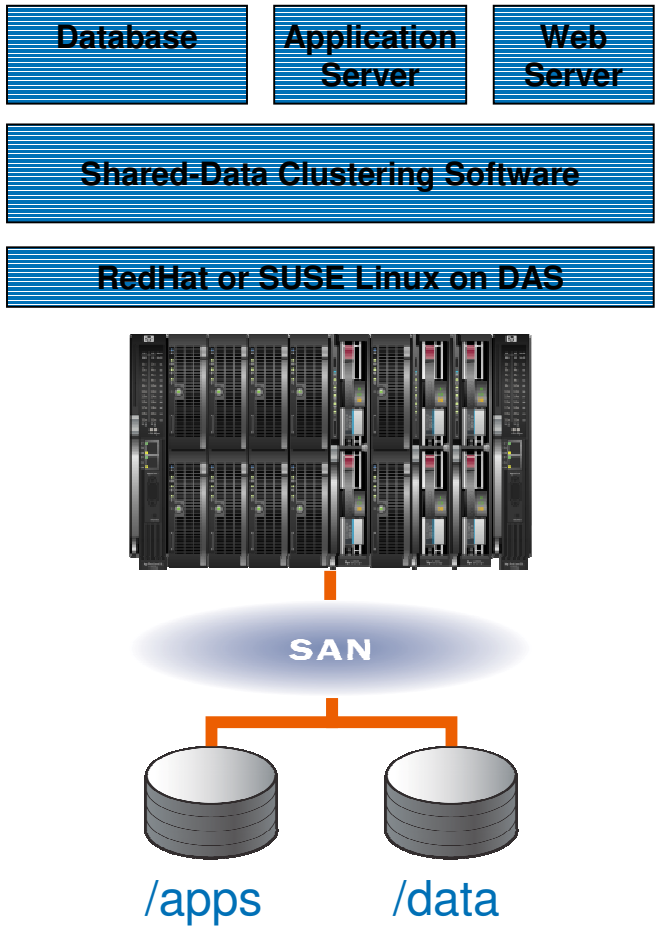
Storage models: iSCSI

- Benefits of back end SAN
- iSCSI router connects network to FC
  - No per-blade adapter costs
  - Longer distance interconnects
  - Redundancy options
  - Long distance bridge via SAN or network
- Block based attributes of SAN
- Moderate to high performance
  - Driver performance evolving
  - Accelerated NICs evolving
  - iSCSI or FCIP protocol options



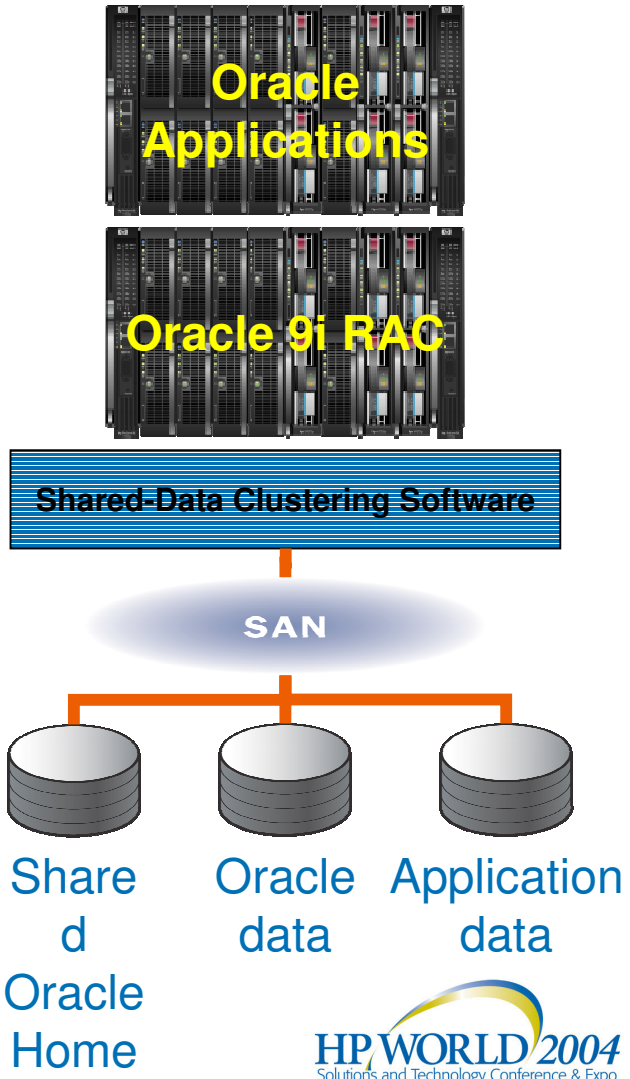
Shared data: Cluster File Systems

- File lock mgmt for block storage
- Asymmetric & symmetric models
  - Metadata system (Lustre, SANergy)
  - Distributed (PolyServe, RedHat GFS)
  - Full speed SAN access possible in both
- Shared access to data files
  - Servers mount /data filesystem
  - Full concurrent read/write access
- Shared access to applications
  - Servers mount /apps filesystem
  - Install, update, patch once
  - N+1 failover, dynamic capacity mgmt



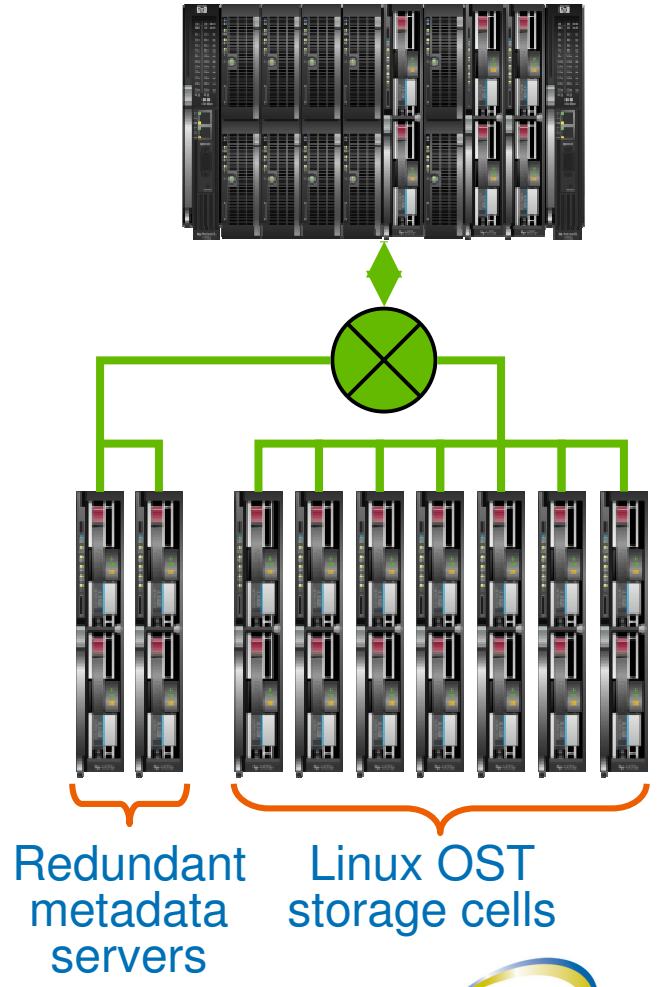
Does Oracle need another CFS?

- Shared Oracle home today
  - Install/upgrade/patch once
- N+1 failover, expansion
  - Replace / add blades by mount
- Uses conventional file systems
  - Standard backup tools
  - Access to data by other applications
- Support for file-based applications
  - Shared data access in SAN



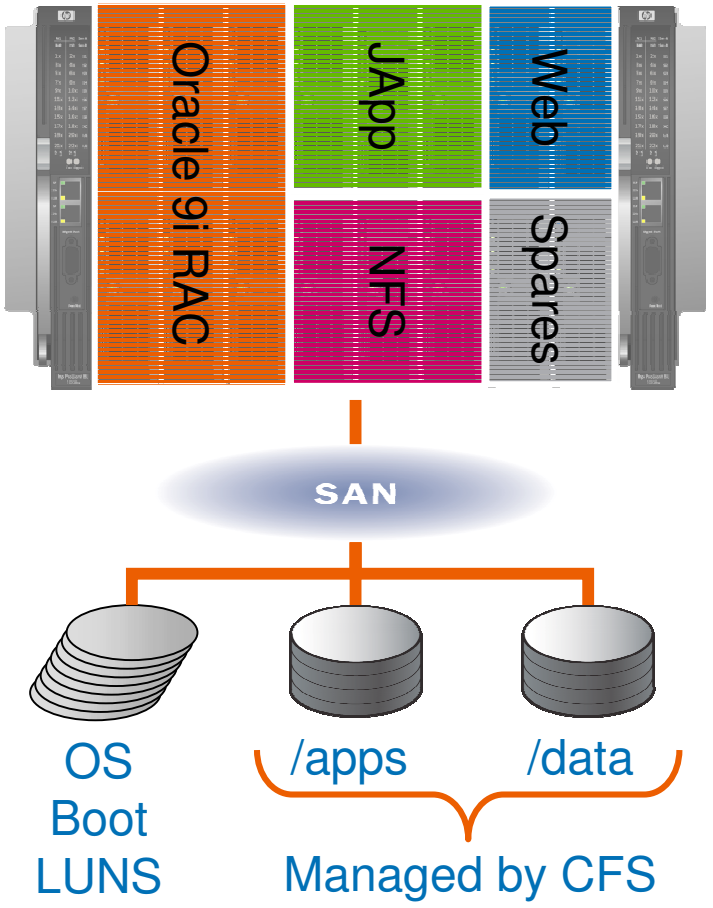
Lustre: a scalable object store

- Open source project
  - Hosted by Cluster File Systems, Inc
  - Scientific computing focus to date
- NAS-like object storage “cells”
- Metadata broker for access
- High performance via parallelism
- Fast interconnect options
- Similar to video server model



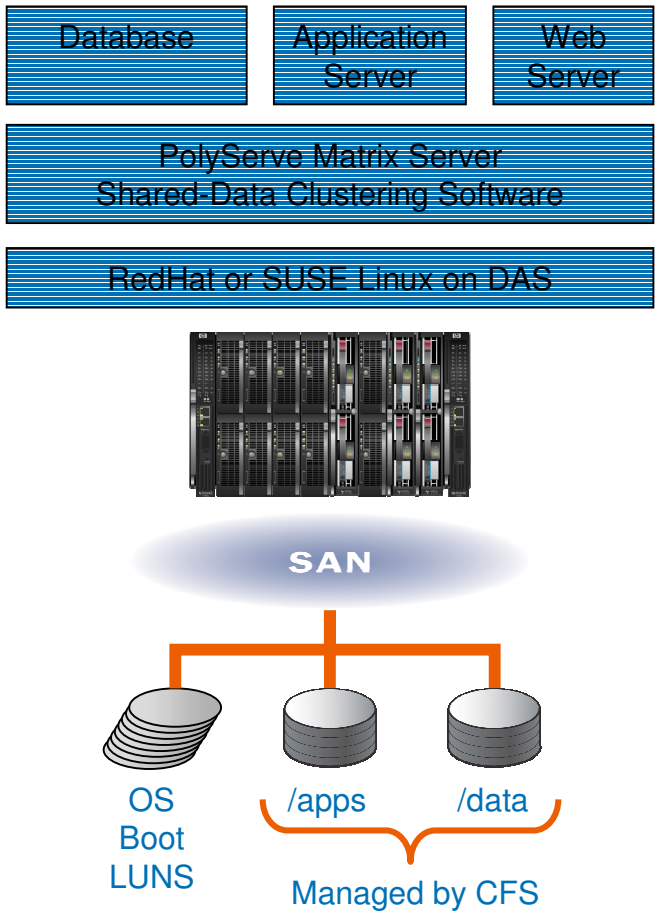
Diskless blades: Boot from SAN

- Nothing loaded on blade DAS
  - No imaging or scripted installations
  - You are what you boot
- Rolling upgrade
  - Build new OS image on boot LUN
  - Reboot server to new OS level
- N+1 failover from spare pool
  - Requires triggers and scripts
- Workload rebalancing
  - Remove blades from underused pool
  - Add to oversubscribed pool
  - Time of day role changes



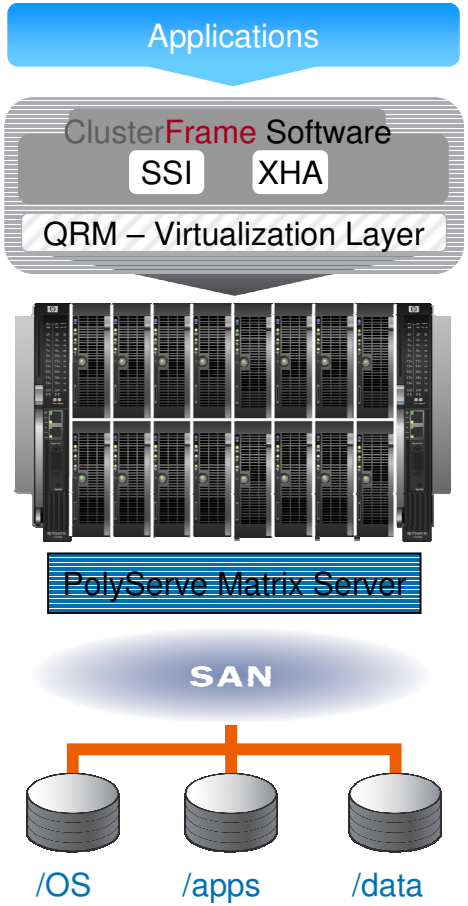
CFS enhanced functions (PolyServe)

- Integrated high-availability
  - Hardware health monitoring
  - Software health monitoring
  - Configurable failover policies
- Single point of administration
  - Central console for unified admin
  - Multi-server view
- Fully scriptable CLI



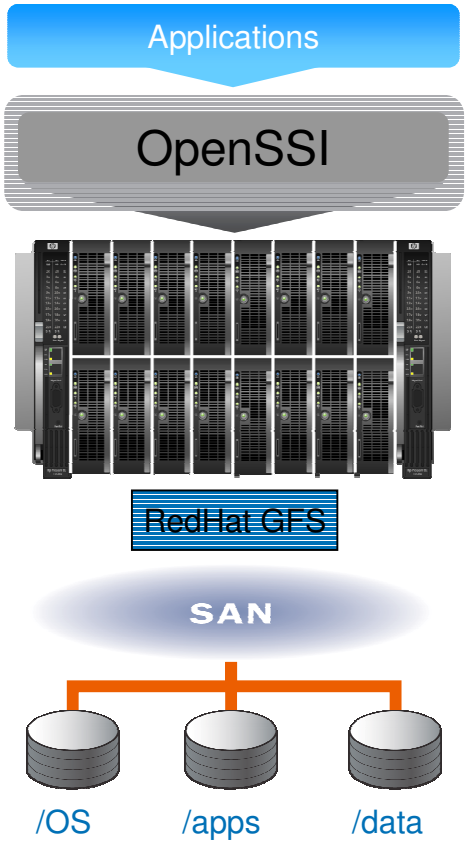
Virtualized commercial cluster example

- Qlusters ClusterFrame QRM
  - Policy-based blade provisioning & mgmt
  - Capacity on Demand synergy
- Qlusters ClusterFrame SSI
  - Load leveling among physical servers
- Qlusters ClusterFrame XHA integration
  - Fast failover of stateful applications
- PolyServe Matrix Server
  - Cluster File System for data and flat files
- HP p-Class blade servers
  - 16-32 CPUs in 6U of rack space
- Implementation Services (recommended)



Open source cluster example

- OpenSSI
  - VAXcluster / TruCluster model
  - Physical servers collaborate as one system
  - Homogeneous view versus “home node” view
  - Manual expansion by booting into cluster
  - Capacity on Demand purchasing synergy
- RedHat GFS
  - Cluster File System for data and flat files
  - Scale-out to hundreds of nodes
- HP p-Class blade servers
  - 16-32 CPUs in 6U of rack space
- Implementation Services recommended



Credits & links

- www.beowulf.org
- www.lustre.org
- www.openmosix.org
- www.openssi.org
- www.polyserve.com/hp
- www.qlusters.com
- www.redhat.com
  - /software/rha/gfs/
- www.vmware.com
- www.hp.com
  - /go/blades
  - /go/htpc
  - /go/linux
  - /solutions/activeanswers
  - /solutions/highavailability/oracle

Summary

Linux blade system clusters offer:

- A low-cost, industry-standard alternative to SMP systems
- Independent optimization of compute & storage modules
- Modular upgrade and very large scale-out options
- N+1 failover options for cost-effective availability
- Solutions tailored for popular SMP applications
- Commercial and open source software stacks

A high-growth platform for IT innovation

HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:

interex

shared knowledge • shared power

encompass

AN HP USER GROUP

RECOMMENDED TRAINING VENUE FOR THE

HP Certified Professional