

# GRID @ Sun

**Richard Croucher,  
Technology Director/Chief Architect  
Sun Microsystems Inc**

# Agenda

- Why Grid
- Grid Computing in Sun
- Sun Grid Utility
- Evolution of Grid Computing for the Enterprise

# Why Grid now?

- Cost economics of building high end SMP compared with single-board systems
  - > Increasing gap between memory speed and processor speed
  - > Move to multi-core and Chip multi-threading enables extremely powerful single board computers to be produced
- Grid has spent the last decade evolving and maturing in academia and with the rocket scientists
- Many enterprises successfully explored Cycle scavenging Grids – need more predictable service now
- Applications are now emerging which can only reach their required performance on a Grid

# Sun and Grids

- Sun established early lead for Grids – coupling SPARC SMP servers
  - > Many Solaris Grids installed and operational at customers
  - > Test bed for high speed clustering – SCI, MyraNet, Sun Fire Interconnect
  - > Internal Grid User with 14,000 “CPUs” deployed in ranches
- Member of Global Grid Forum and Enterprise Grid Alliance
- One of the largest install bases of DRM with N1 Sun Grid Engine
- Current x86 price/performance lead has seen massive move to Linux
  - > Sun delivered multiple, very large Linux, x86 Grids
  - > Sun is shipping highest volumes of Opteron 2-4 CPU Servers
  - > Sun is first to market with 4 socket dual-core Opteron server
  - > Solaris 10 x86 launched – All the advantages of Solaris on commodity servers
- Extensive experience building and deploying Data Centers to achieve 99.99% reliability
  - > Sun now applying, same design patterns and component based methodology to deliver Enterprise Grids designed to provide level of reliability necessary for compliance
  - > Featuring High performant LAN design, Storage, NAS clusters, InfiniBand, Provisioning and Management.

# N1 Grid

- Grid Computing is a core component of Sun's N1 Grid Vision
- Grid Computing focuses on the computational environment and workload management in the data center, and is the subject of this presentation
- N1 Grid is Sun's vision, architecture, products, and services for optimizing network computing
- N1 Grid System provides all of the core services for establishing, partitioning, provisioning, and managing grids

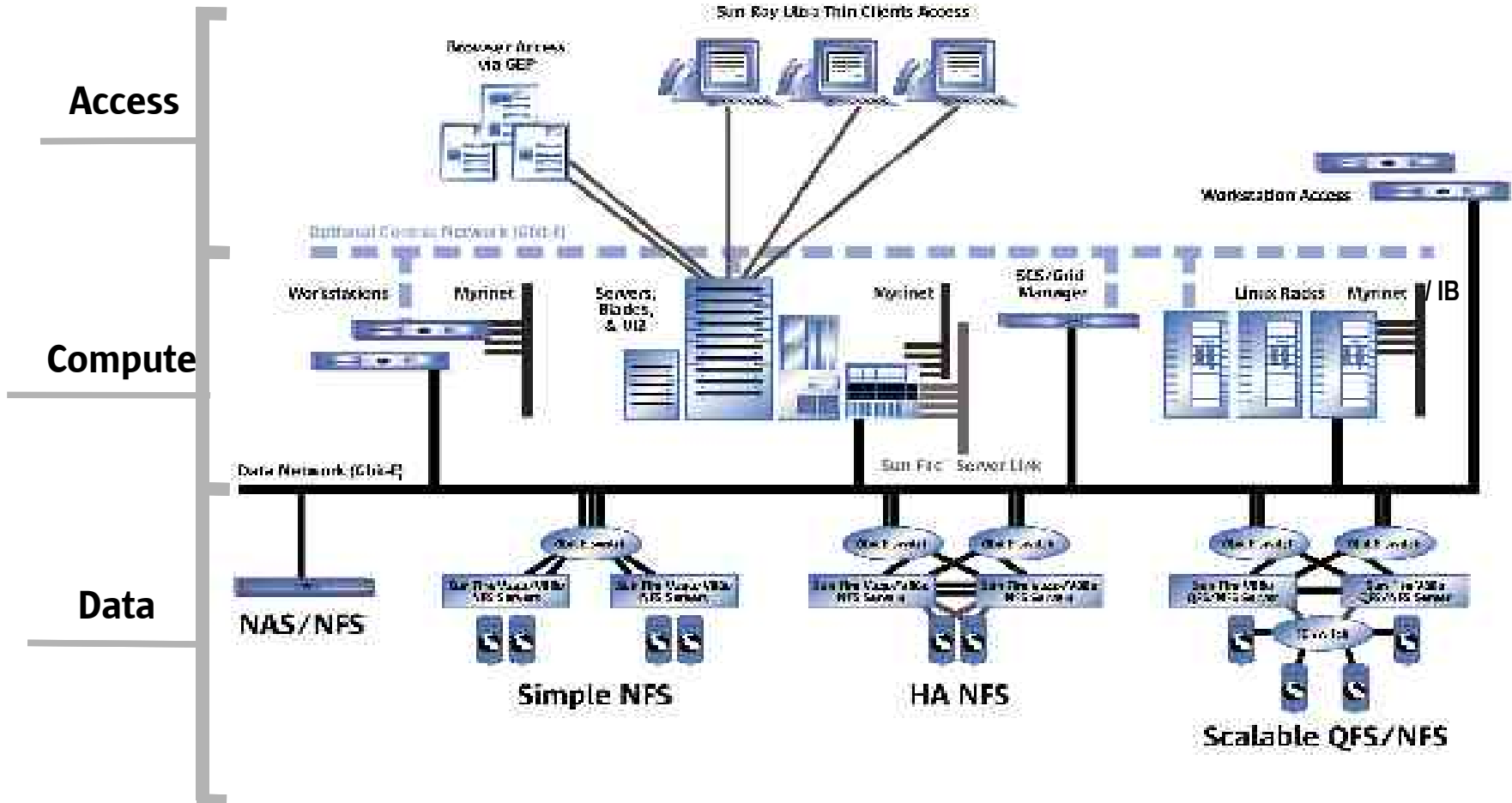


**N1 Open and Integratable**

**- published API's enable integration of 3<sup>rd</sup> party HW and Software**

# Grid Reference Architectures

## Proven and Repeatable Solutions



# Sun Grid “Factory” Customer Ready Systems – Scotland



**e.g. European Customer  
7 pods @ 128 nodes  
Sun designed and supplied  
network**



# When to Grid?

- Requires more effort to make applications Grid enabled
  - > Always simpler to utilise the traditional single memory SMP model
- Design new applications so they can be Grid'ed
  - > Leverage Jini
- Select your Grid middleware and support toolsets
- Ask your ISV's what their plans are?
- Evaluate existing application portfolio for Grid candidates
  - > Web services
  - > Horizontally scalable – survive node failures
- Look at adding InfiniBand into your infrastructure
- Access the Sun Grid Utility



# Sun Grid Utility

- New business model for Grid Computing based on traditional utilities model
- Pay for what you use
- Sun caused a storm when it announced at \$1 per CPU-Hour
- Predicted to be the way most corporations will buy compute power in the future
- Solves the customer problem for deploying ever hotter CPU's

# Sun Grid Utility Baseline offering – Technical Details

- Initial standardized configurations:
  - > Sun V20z System with AMD Opteron processors
    - > 4GB RAM per CPU
  - > Gigabit Ethernet compute & management network
  - > Managed infrastructure
- Software licenses (including support)
  - > Sun Solaris™ 10 Operating System
  - > Sun N1 Grid Engine 6
- InfiniBand option available



# Sun's Grid Utility

- Shared Grid Computing Farm
- Provided as a Managed Grid Computing Service
- Utility, pay-as-you consume charging
- Designed to be commercially advantageous to most customers
- Creating a new market and evolving the business models
- Retail and Commercial Service Models

# Sun Retail Grid Utility

Internet portal access, Solaris 10, N1-SGE, pay-per-job, pay in advance with Credit card, shared Grid environment.

- > Upload job, wait for notification of job completion, download output
- > Available to the public
  - > Regulatory issues restrict initially to US

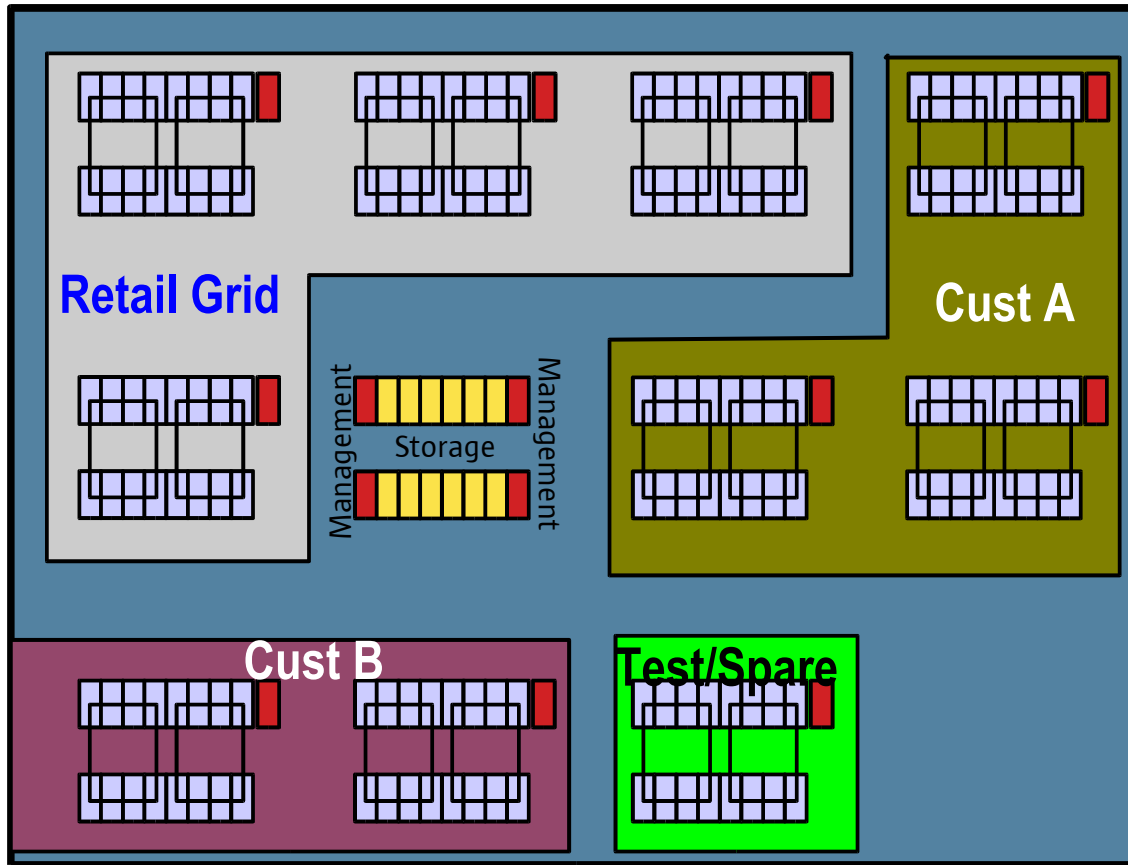
# Sun Commercial Grid Utility

- Systems allocated exclusively to you during your rental period – private Network acts as extension to your existing systems
- Offering details
  - > 4 hour reservation periods – pay for it when you need it
  - > Provisioned as you specify then scrubbed when done; No persistent state maintained on systems
  - > Gigabit Ethernet standard; InfiniBand available
  - > Customers “Golden image” deployed for their reservation
  - > Solaris 10 and N1-SGE included within utility pricing
- Can flex capacity requirement at short notice
  - > % spare capacity kept available
  - > Automated provisioning of new equipment accelerates massive expansion

# Sun Grid Utility Benefits

- Flexibility
  - > Pay for what you need
  - > Compute services on demand
- Enable consolidation of existing Grid silo's
- Avoid data center re-planning
- Improved standardization
- Partnership with Sun
  - > Implicit technology refresh
  - > Technology based solutions
- Managed service
- Customer choice – hardware/ software ..

# Sun Grid Utility Data Centre

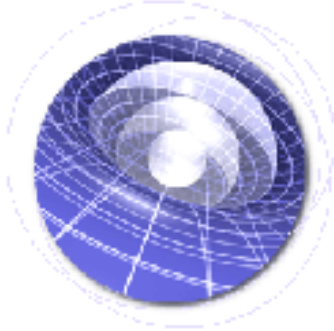


**Scalable to >20,000 CPU's in a single Grid Ethernet Cluster**  
**Do you have 50,000 sq. feet of DC space spare?**

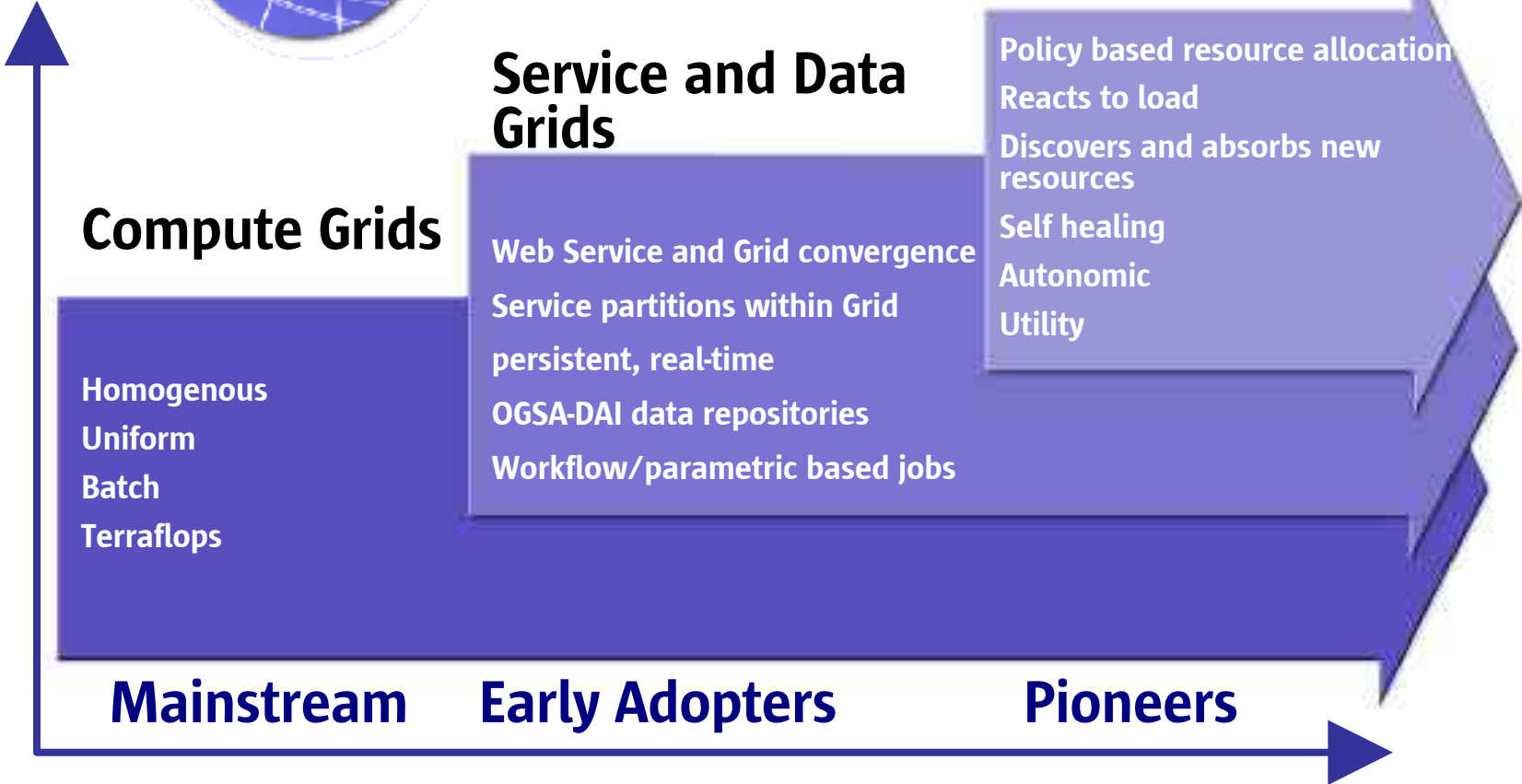
# Grid Evolution and futures



# Grid Maturity Levels

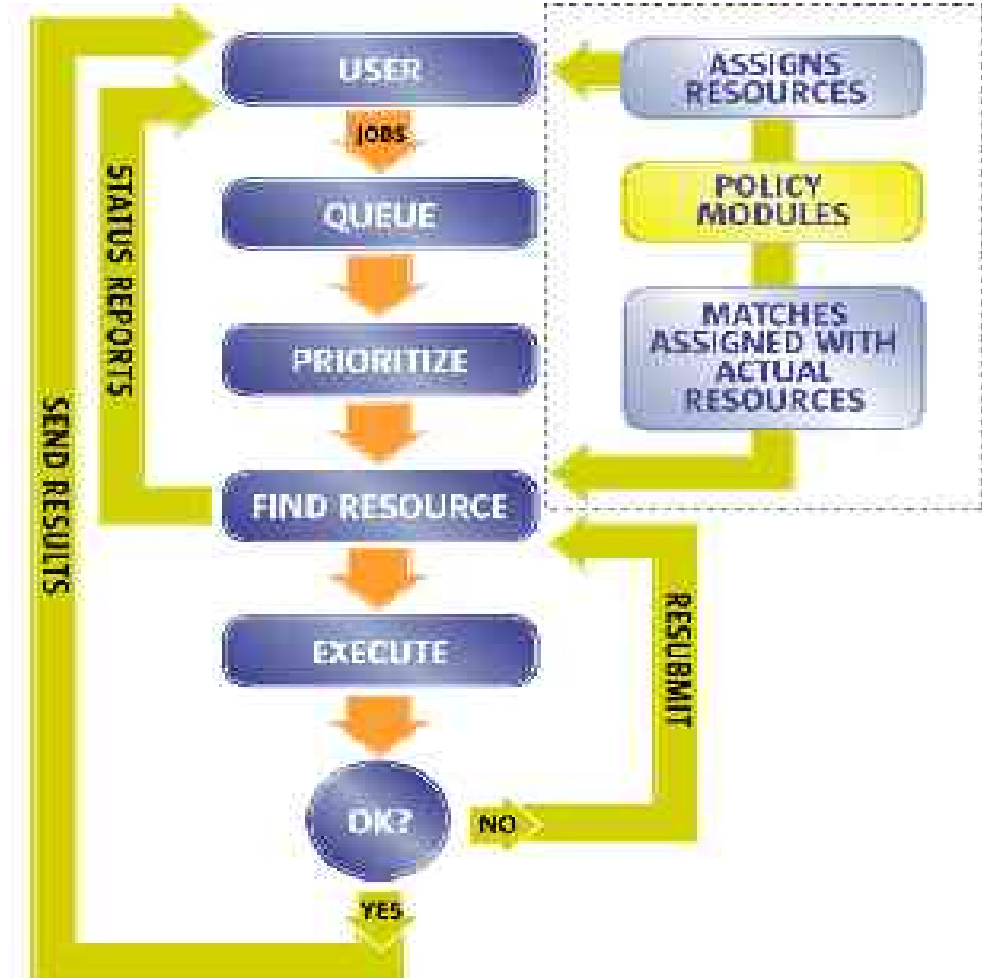


Terminology becoming standardised in industry – Bloor, IDC, TAB



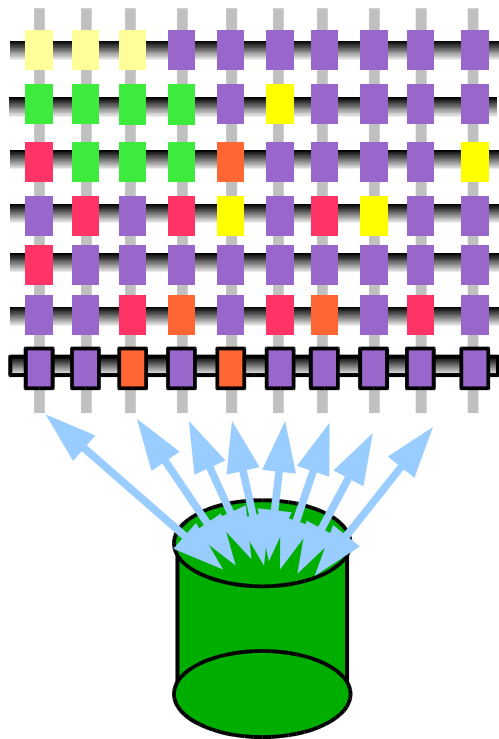
# Compute Grids

- Harness low cost hardware to solve parallizable problems
- Homogenous, fixed sized nodes, typically 1-2 CPU
  - > 32-10,000 nodes
  - > Rendezvous after each parallel job
- Predominately batch workloads. Nodes wait for their next job.
- Terraflops performance
- Can run on Cycle scavenging Grids
- Often use MPI libraries
- N1 Sun Grid Engine, Condor Platform, DataSynapse



# Service Grid

- Convergence of Compute Grids and Web services
- Common fabric which provides access to interconnect, storage and LAN
- Logically partitioned into smaller 4-128 node Grids
- Nodes are assigned into logical service groups, to provide specific services or capability:
  - Specific services e.g. Order Entry, often J2EE based
  - “Sticky” = nodes are allocated to a specific service for long periods
- Common fabric enables provisioning of resources to groups and multi-purpose nodes tier 1,2, or 3
- N1 like, Utility ready
- API's – OSGA, Remote memory access (DAPL), Real-time DRM's , SDP



# Data Grid

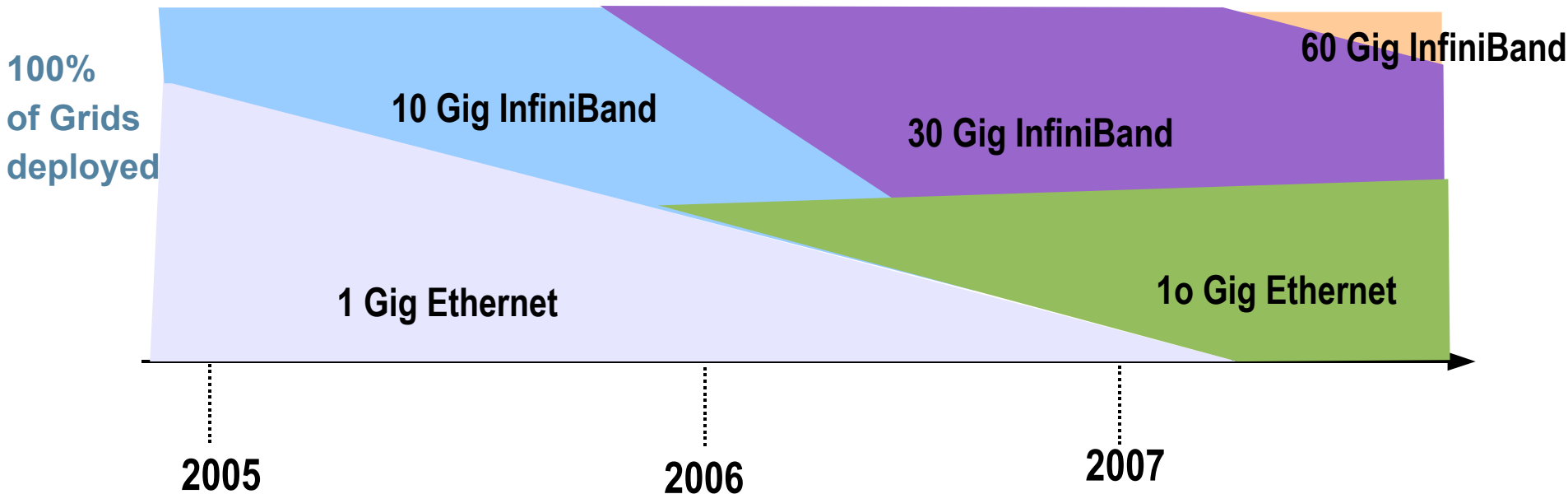
- Access and transformation of data
- OGSA-DAI is driving the reference model
  - > Register existing data sources
    - > Oracle, XML, ObjectStores, CSV files
  - > Made accessible to Grid jobs
    - > Add, extract and search registered repositories
  - > Semantic webs and ontologies emerging where required
  - > Driven by:
    - > Life Sciences (Genome, Protein Databases, Engineering parts databases)
    - > Big science experiments – Cern LHC, Global warming etc.
    - > Digitization of media
- Enterprises typically carry out more data manipulation than algorithmic computing typified by first generation Grids
- First to exploit will gain competitive advantage

# Common fabric enables the Service Grid

- A common fabric to which all systems attach, guarantees connectivity and permits services to be moved to the resources they need
- Service Grid enabled applications can exploit efficient Remote Memory Access
- InfiniBand provides a common backplane for the service grid, with a Powerful transport engine to eliminate software stack overhead. It provides:
  - > Cluster interconnects
  - > LAN attachment
  - > SAN attachment
- Now less than £1000 per attached node (HCA + switch port)
- Any node can provide Tier 1,2, or 3 level services
- Backwards compatible to enable support for non-Grid enabled applications via existing TCP/IP, NFS and SCSI driver stacks
- Ethernet will continue to be used for the Compute only Grids

# Anticipated Ethernet/InfiniBand distribution in Grids

1Gig Ethernet will continue to dominate cost-sensitive compute grids. Move to Service Grids and the need for flexible infrastructure will drive InfiniBand. InfiniBand is targeted as a cluster interconnect, not a LAN. It will continue to have a performance lead over Ethernet as a Grid interconnect.



Cost per attached node is critical in Grids.

10 Gig Ethernet is currently 2x cost of the 2-CPU “sweetpot” Grid compute node. Cost needs to drop to < 50% of the Node cost before widespread deployment will occur. This is not predicted until late 2006/7. After which it will replace 1G as the lowest cost network attach, primarily as a result of integration into motherboard I/O chips

Many API's common to both e.g. DAPL, iSER/iWARP, SDP

# Service Oriented Architecture and Service Grids

SOA - Principles and practices for designing shared, reusable distributed services

## Attributes:

- Separation of service interfaces from underlying implementation (loose coupling)
- Promotes service reuse through discoverable and self-describing services
- Services are coarse-grained, composable and rely on a standards based infrastructure

## Implications:

- Look to SOA to provide agility and flexibility – deliver services to market quicker
- Components need to be highly scalable
- Use a flexible Grid infrastructure to underpin SOA
- Virtualizes the resource – allocate them to where they are required

# Next Generation Grid Environment

- Needed to provide a common platform to “Grid enable” applications
- “virtualises” the Grid into a simple execution environment.
- Possible choices:
  - > Single System Image
  - > Open Grid Services Architecture (OGSA)
  - > Vendor API – e.g Platform, DataSynapse
  - > Jini
- The winner will be the one which pulls in the ISVs and developers



# Single System Image

- Extend a single OS to span the Grid Cluster
- InfiniBand RDMA is an enabler
- Several Linux based activities
  - > Mosix/Qclusters, Scyld-Beowulf, OpenSSI project
- Two major issues to contend with:
  - > x86 compute nodes trade-off reliability for performance – typical failure rate is 2-12% in any Grid deployment
  - > Distributed systems fail in too many arcane ways to make tightly coupled cluster reliable enough for real world computing
  - > Not reliable enough for most Enterprise computing
- Sun delivered SSI with Solaris Multi Computer project
  - > Determined not to launch commercially

# Open Grid Services Architecture

- Defined by Global Grid Forum (GGF) <http://www.ggf.org>
- Implemented by Globus - <http://www.globus.org>
- Uses XML based Web Service Definition Language (WSDL) transported via SOAP and published using UDDI
- Constantly changing e.g toolkits v2.2,v3,v4
- High interest from academics, little in Commercial sector
- Potential long term play, but any current work will be a throw-away
- Concern is whether it runs the risk of collapsing under it's own weight of “standardisation”

# Extending a DRM to the Service Grid?

- Distributed Resource Managers have been designed for batch
- Enhancements made by both Platform and Data Synapse allow pseudo real-time like capabilities
- Both driving their own agenda's with proprietary API's
- Are gaining ISV traction and good expedient option for short term
  - > Expect to have to replace as standards solidify

# Jini Grid

- Jini provides an intrinsic solution to the distributed computing problem
  - > Handles classic issues around reliability and resiliency
  - > Leverages but not limited to Java. Can attach a Jini agent to allow inclusion of other code modules
  - > Java Spaces provides a distributed object store
  - > Service centricity aligns well with SOA – better than J2EE
- Stable, well established API and extensible services
- Several major initiatives around Jini Grid computing
  - > Open Source available
- Several JavaSpace implementations offering persistent, clustering, improved performance, InfiniBand support coming
- Sun RIO project adds Java Service Beans, QOS, Dynamic Service creation
  - > Provides a robust environment proven in mission critical application deployments
- <http://www.jini.org>



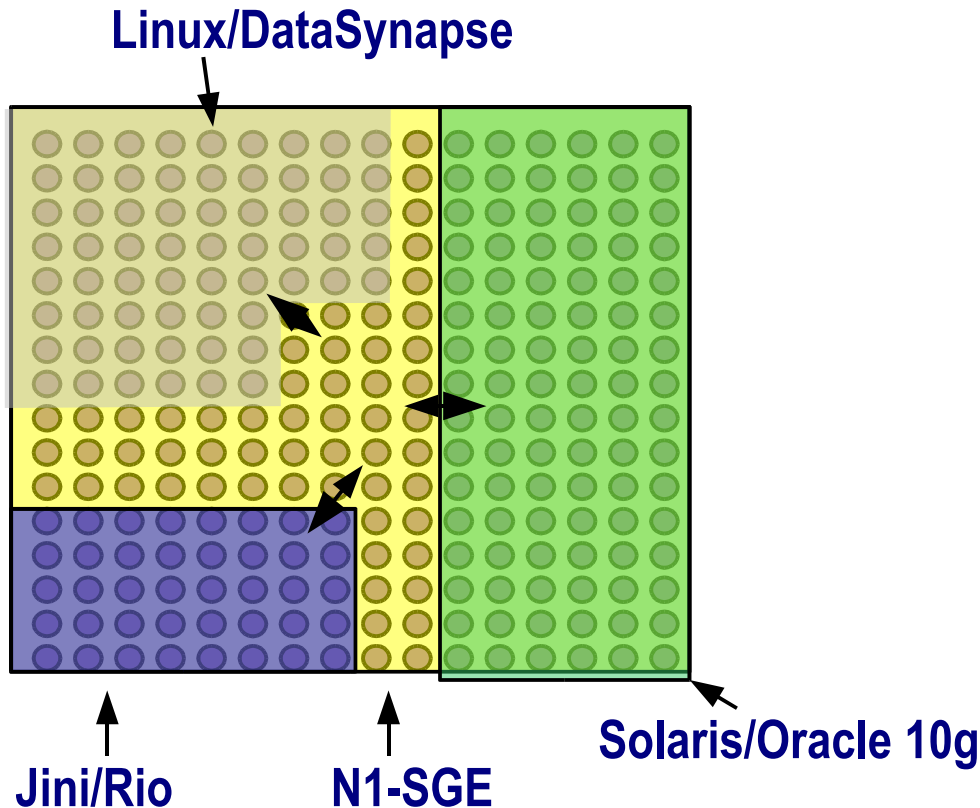
# Why Jini?

- Jini network technology is an open software architecture that enables developers to create network-centric applications that are highly adaptive to change.
- By design, Jini network technology can be used to build adaptive networks that are scalable, evolvable, and flexible, as typically required in dynamic computing environments

## What does Jini Do?

- Tracks services in dynamic environments
- Identifies services among those currently available that can satisfy client needs
- Assembles services into live system at runtime
- Coordinates interaction of distributed components
- Provides solutions for many of the issues around distributed computing

# Grid Consolidation into Service Partitions



**Service Partitions each with different Workload Scheduling schemes**

**Each partition implements own policies for workload management**

**Policy based Resource Mngt, allocates resources into a domain**

**Each partition can support different OS build and different Middleware**

# Service Partitions

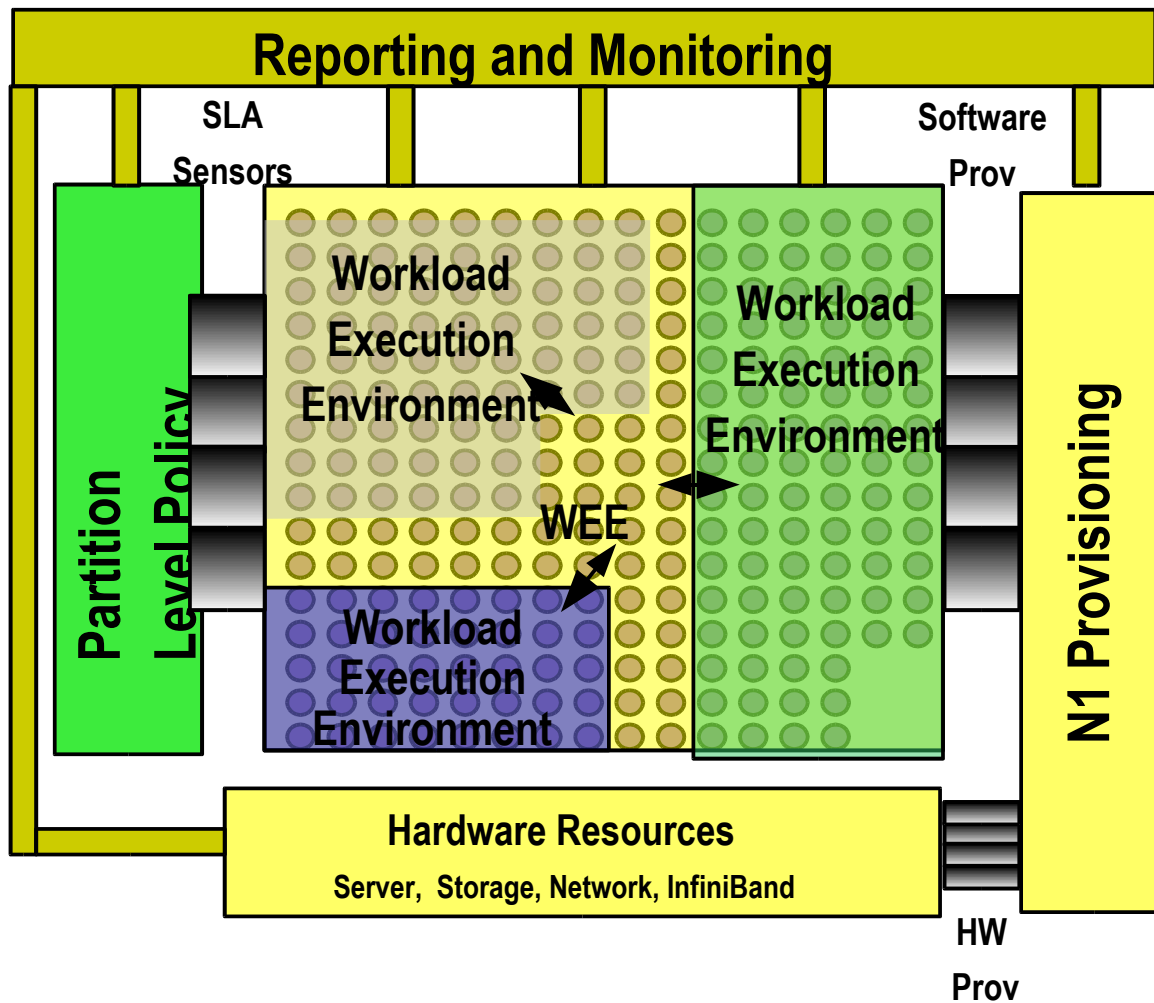
- Personalization of a compute node to provide a specific service
- Deploys the Application + OS stack specific to a Service Partition
- A common Grid HW infrastructure enables any Node to take on the necessary personality to be part of a Service domain
  - > Need tools to rapidly deploy Application Stacks – N1 SPS
  - > Need tools to rapidly deploy OS builds – N1 System Manager
  - > Need tools to monitor and manage HW – N1 System Manager
  - > Need a policy/ rules engine to move resources to/from Service Partitions - Enigmatic EMS

# Grid Service Partitions

- Need to be hierarchical
  - > Support both organizational and “Application domains”
- Resourcing changes to a partition take place in hours/minutes not seconds.
- Could be accelerated by booting Solaris Zones – but is not meant to be changing continuously
- Hysteresis and dampening to prevent thrashing
- Allow for Manually initiated allocation in expectation of an impending event



# N1 Grid – Open and Integrateable



# Integration points and API's

- Service partitions wrap around specific Grid execution environments – can be done to existing binaries. Behave as if they are running on a smaller physical Grid
- Dynamic addition of resources assumes the App stack has some mechanism to announce their availability. Most major Grid apps (N1-SGE, DataSynapse, Platform, Jini/RIO) all provide this
- Dynamic removal of resources assumes the App stack can cope with a seemingly failed hardware node. Most have been designed to do this, but demand varying degrees of job re-running
- Service Level Monitoring could make use of existing HA-Agentry for application specific probes

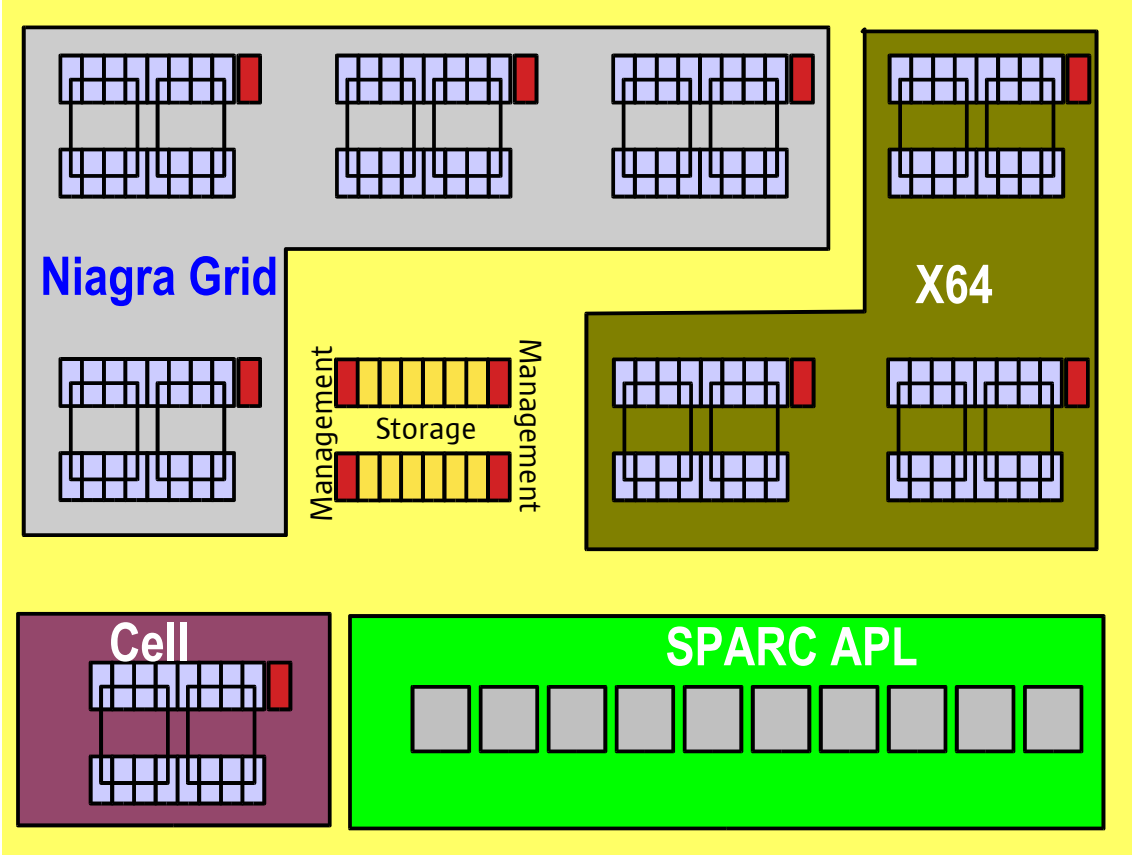
# Automate the orchestration

- Requires sensors to determine compliance with service level metrics – Response time, load level
- Requires Rules and policies to determine where to allocate resources
- Futures is to utilize constraint based algorithms to optimally use a finite pool of resources

# New Hardware architectures will drive specialization

- New developments in computer architectures are predicted to lead to a diversification of architectures
- Power and Cooling starting to impact new designs
- X64 – commodity player. Good compute and floating point, but very power hungry. Course grained parallelism.
- SPARC (Niagra) – First Sun CMT, 8 cores, 32 threads optimized for data manipulation. Power miser
- Cell – Sony/Toshiba/IBM – designed for PS3, Grid on a Chip design, optimized for floating point. Course grained parallelism.
- SPARC 64V – Sun/FJ APL - Big memory footprint >1TB, fine grained parallelism, traditional applications

# Next Generation Data Centre



Individual applications span the physical Grid's to obtain the right resources

# Summary

- Grid Computing becoming mainstream
- Make Sun your Grid partner of choice
  - > 22 year UNIX heritage
  - > Continuous innovation and openness
  - > Long history building and deploying Grids
  - > Leading price/performance with Sun V20z/V40z
  - > Innovating with Chip Multithreading
  - > Utility purchasing option
    - > Retail, Commercial
    - > Service Delivery Centres now open in London, New York and Virginia
      - More to follow



**[Richard.Croucher@sun.com](mailto:Richard.Croucher@sun.com)**