Grid Infrastructure: Creating the Utility Data Center

Dr. Carl Kesselman
Chief Scientist, Univa Corporation

April 2006

What is Grid?

- A large-scale, geographically distributed, hardware and software infra-structure composed of heterogeneous networked resources owned and shared by multiple administrative organizations which are coordinated to provide transparent, dependable, pervasive and consistent computing support to a wide range of applications.

"Grid Characteristics and Uses: a Grid Definition"

- A Grid is a system that:
  - Coordinates resources that are not subject to centralized control...
  - ...using standard, open, general-purpose protocols and interfaces...
  - ...to deliver nontrivial qualities of service.

"What is the Grid? A Three Point Checklist"

Do not focus on ***what***, but instead on ***why***...

Why Grid? Driven by Today's IT Reality:

- Moving to distributed, collaborative projects
- Explosion in enterprise accessible data
- Proliferation of "IT islands"
- Over-provisioning = idle computing resources
- Moving to scale-out architectures, using commodity computer, network and storage resources
- Broad set of applications implemented as inflexible "silos"
- Constantly rising management costs, despite decrease in hardware and application software costs
- Slow response to new application rollouts and faults

Why Grid? Driven by Today's IT Reality:

- Moving to distributed, collaborative projects
- Explosion in enterprise accessible data
- Proliferation of "IT islands"
- Over-provisioning = idle computing resources
- Moving to scale-out architectures, using commodity computer, network and storage resources
- Broad set of applications implemented as inflexible "silos"
- Constantly rising management costs, despite decrease in hardware and application software costs
- Slow response to new application rollouts and faults

Current Grid technologies address some of these…

Why Grid? Driven by Today's IT Reality:

- Moving to distributed, collaborative projects
- Explosion in enterprise accessible data
- Proliferation of "IT islands"
- Over-provisioning = idle computing resources
- Moving to scale-out architectures, using commodity computer, network and storage resources
- Broad set of applications implemented as inflexible "silos"
- Constantly rising management costs, despite decrease in hardware and application software costs
- Slow response to new application rollouts and faults

Current Grid technologies address some of these...

Delivery of IT as a service addresses the rest

The Distributed Computing Hangover

- Scale-out approaches have led to explosive growth in the number and distribution of devices that must be managed
- Cost of hardware and application software is dropping...
- ...but infrastructure management costs are skyrocketing
- Application complexity is high...
- ...but infrastructure complexity is higher

IT is failing to deliver against its mandate of *increasing value at lower cost*

Proliferation of "IT Islands"

Need: Apply resources as required for each application

Need: Share some resources among multiple applications

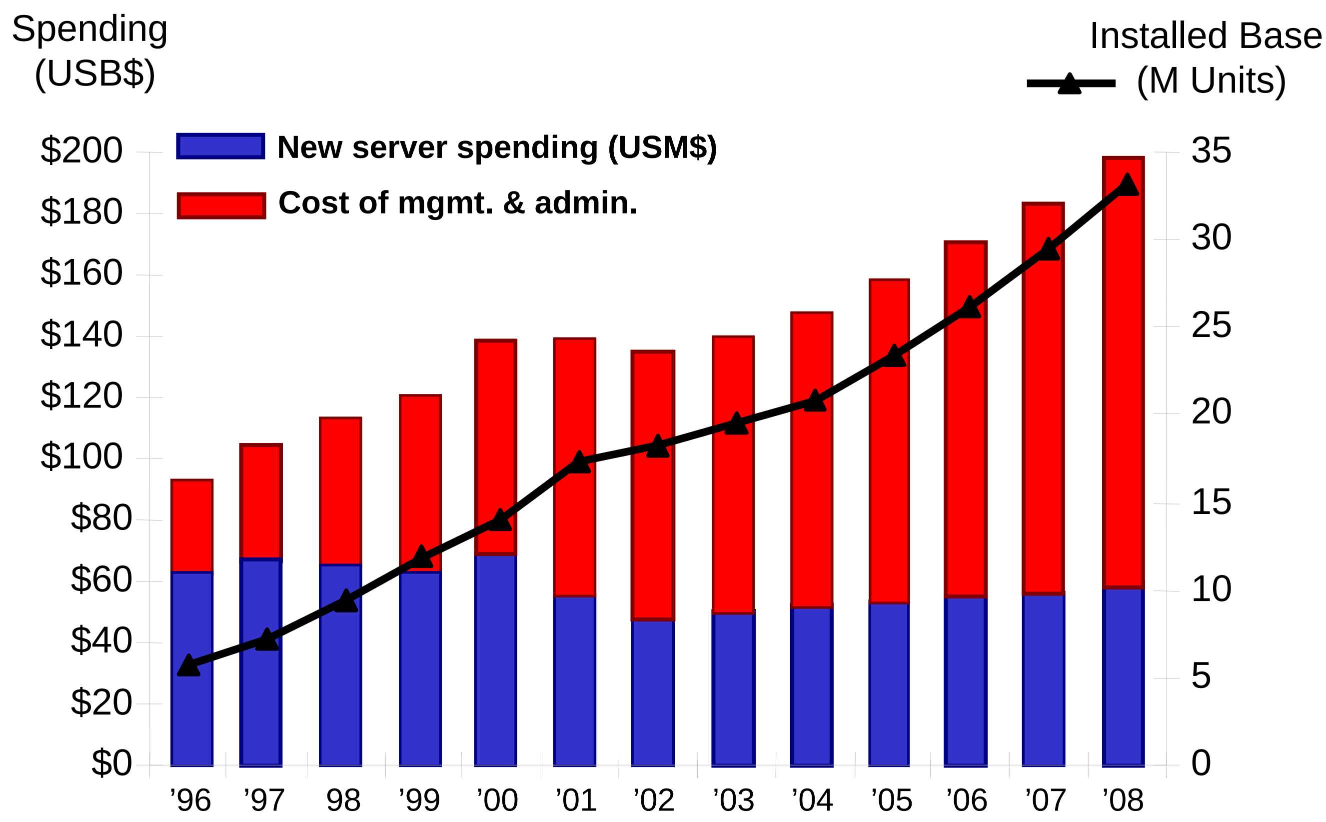
Traditional Approach

Dedicate resources to ensure a quality of service: resource availability, environment configuration, application stability

The Latest "Band Aid"

- Build more/"better" clusters
- Focus on cluster management
- Exacerbated by disaster recovery - more data centers

A Telling Metric:
$7 Management per $1 hardware spend



Source: IDC

Understanding the Management Overhead

	% Of Time
Initial software deployment	18%
Software maintenance, tuning	16%
Migration	14%
Planning for upgrades, expansion, capacity	13%
Upgrades	10%
Server monitoring	8%
Server Maintenance	7%
Reconfiguration	6%
Hardware discovery	4%
Service level management	4%

Focus should be here

Source: IDC

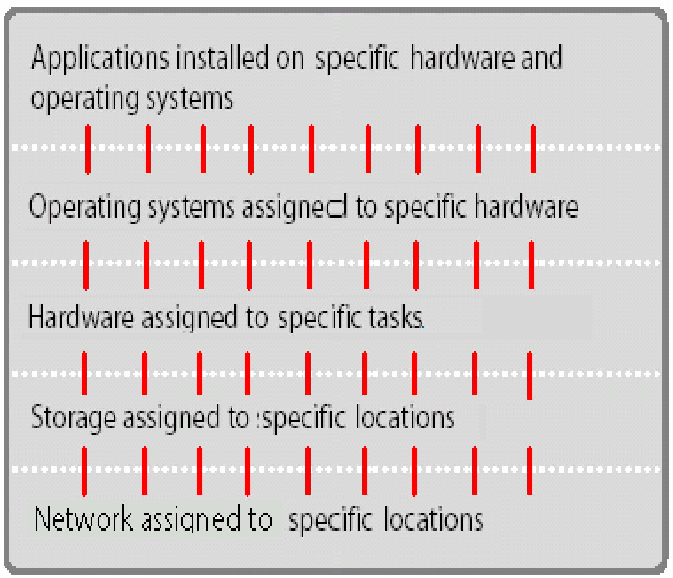
Required: A New Way to Think About Infrastructure Management

- Cluster management techniques will not scale
- Need to move towards scalable, distributed asynchronous approaches
- Need to address data distribution, security, resiliency challenges
- Think about business continuity (value), not disaster recovery (cost)

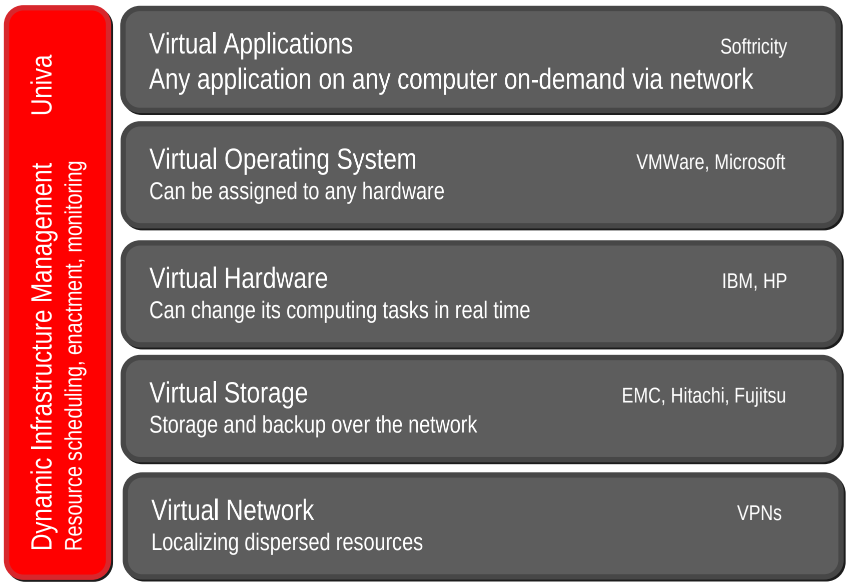
- Allow IT to deliver services to users ***more rapidly***, at ***lower cost***, and at ***higher QoS***
- Grid infrastructure software forms the fabric for service delivery

Traditional vs. Virtual Computing

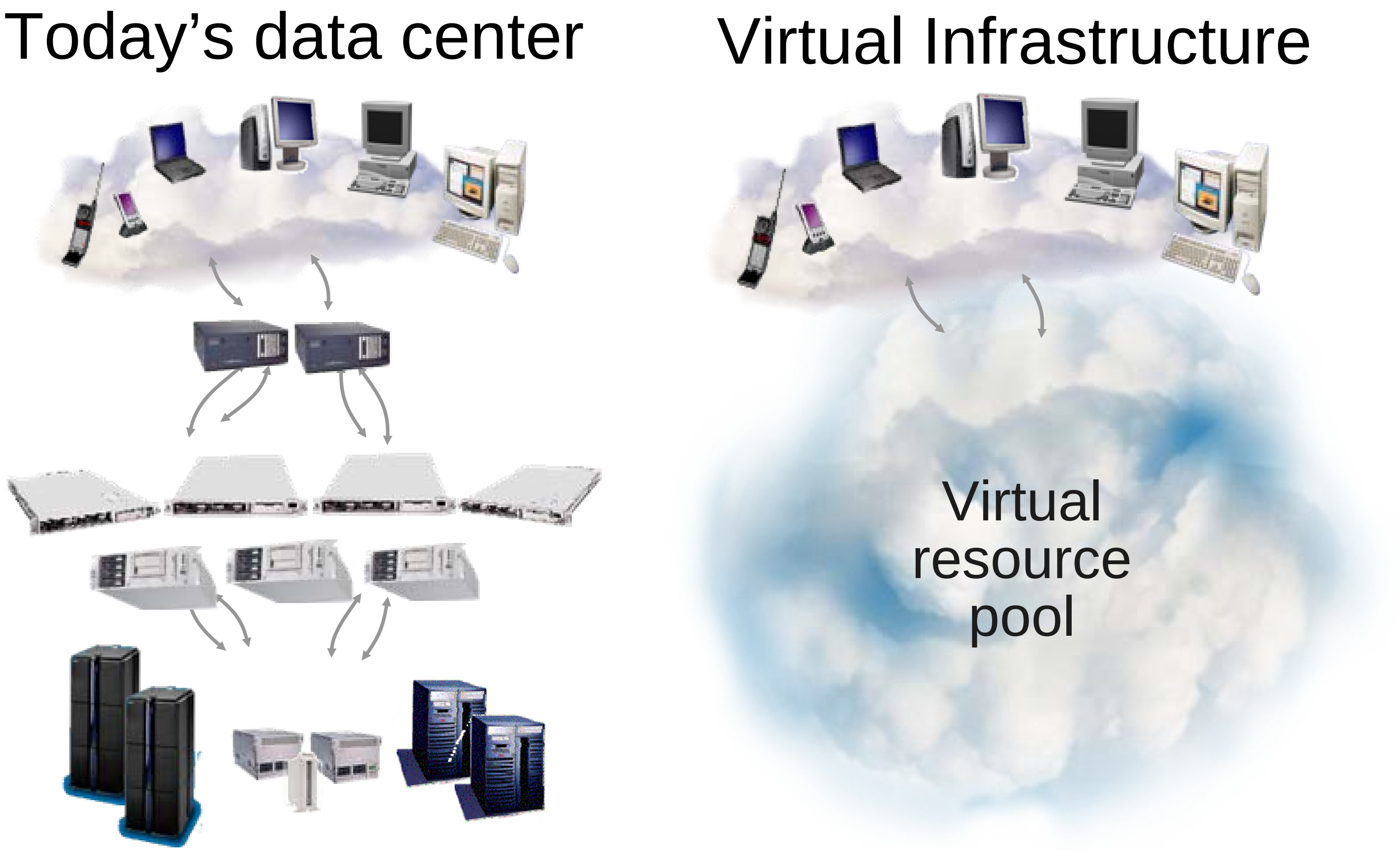
TRADITIONAL



VIRTUAL COMPUTING



Virtualization – driving complexity out



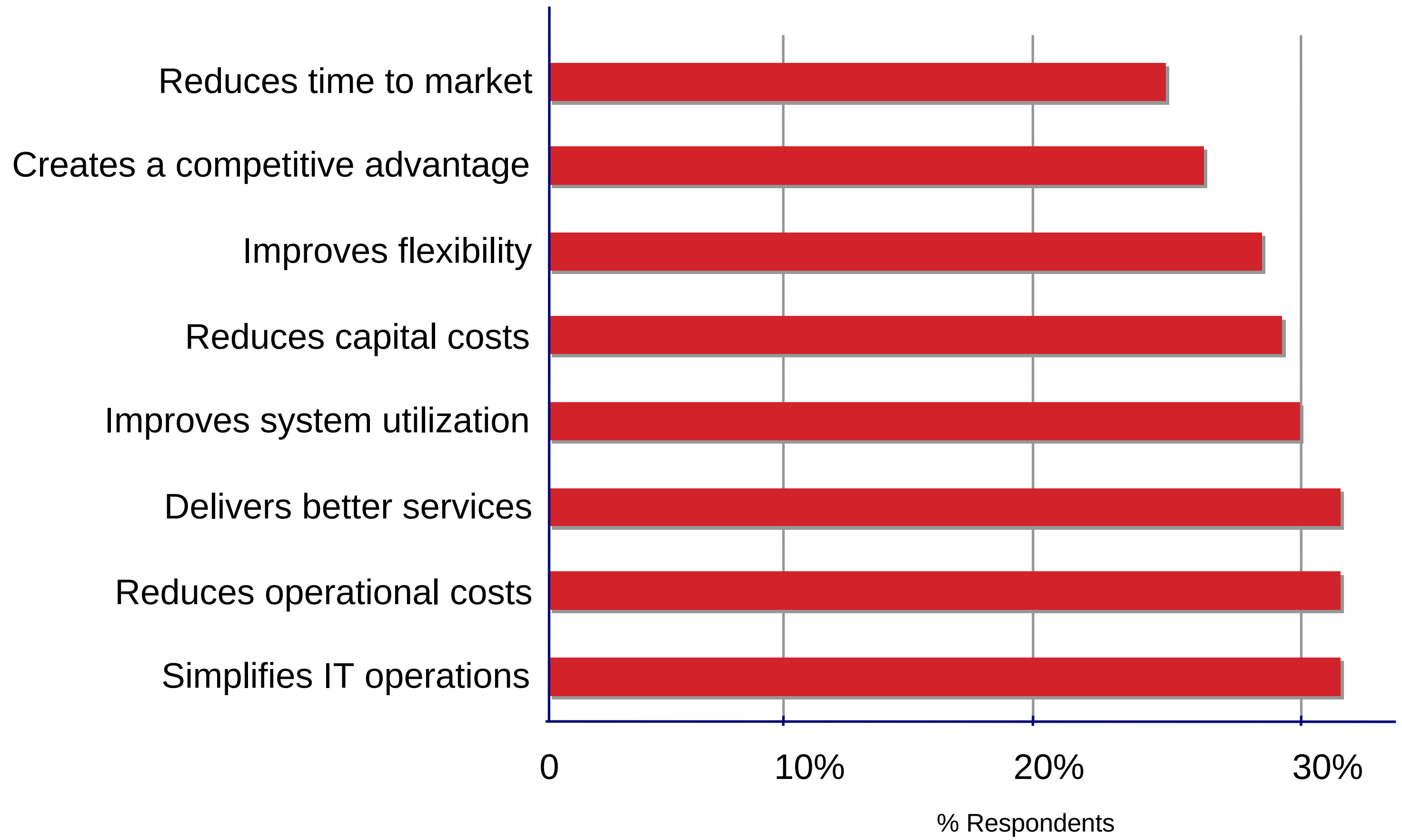
It's All About Service Delivery

IT is about delivering **universal access** to computing services without the common headaches of managing a set of **diverse resources** across a **heterogeneous** and ever-changing set of end-user **demands**.

- Virtualization is the first step
  - Hardware/OS virtualization manages physical device...
  - ...but system still exists...
  - ...and still needs to be managed
- Requirements for management change radically
  - Reduced to machine availability
- Need to isolate the "stateful" tiers in the infrastructure

Univa

Service Delivery: IT Value Proposition



Source: IDC Survey

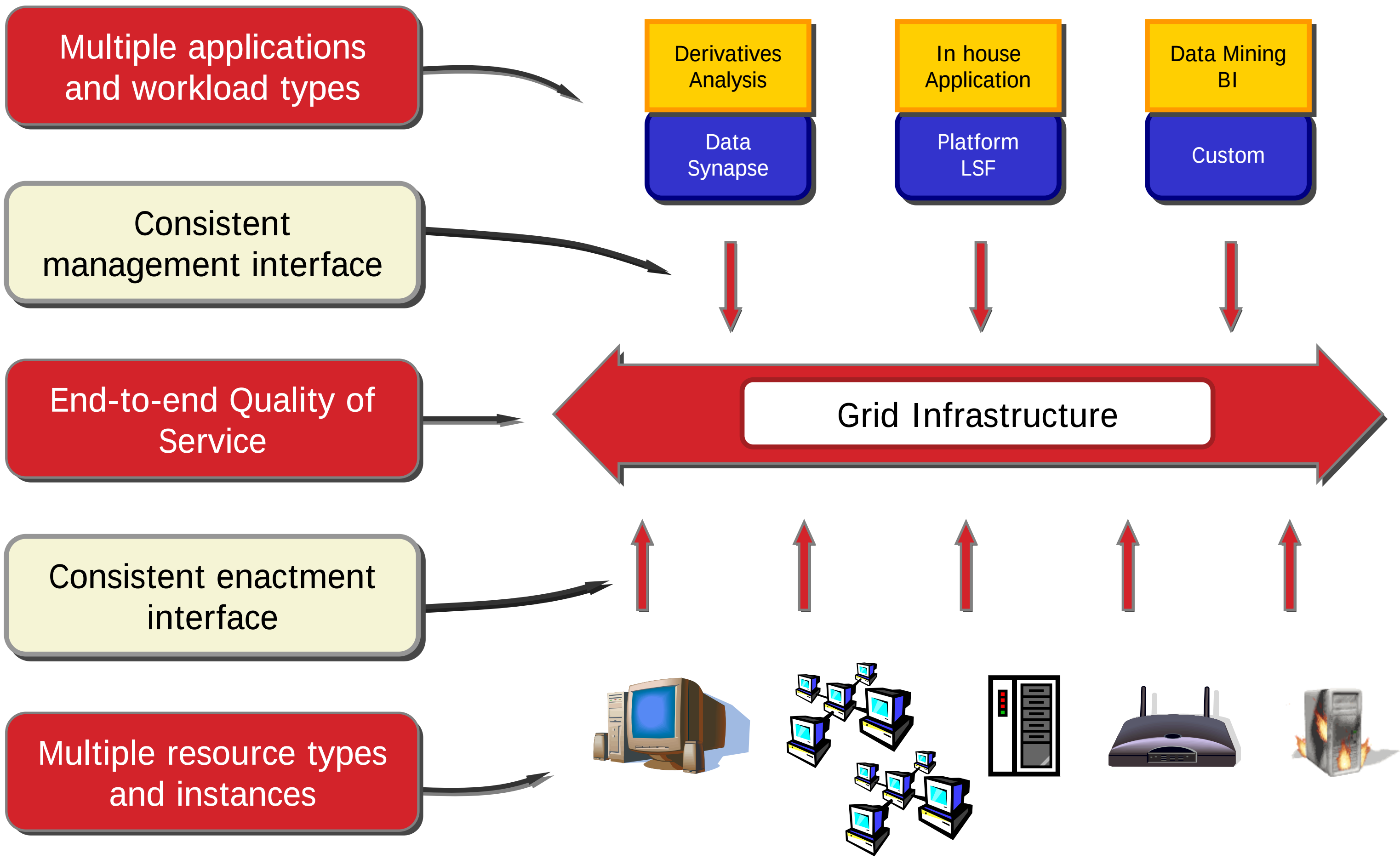
Impact of Service Delivery Approach

Isolating the stateful tiers limits the requirement to scheduling and load balancing

- Schedule globally (with advance reservations, allocations)
- Load balance locally (using really simplified policy)

- **Impact**: No/less need for process or machine migration
- **Impact**: Minimal data required for management/monitoring
- **Impact**: Delivers predictable QoS

Multiple Services Delivered on a Common Infrastructure



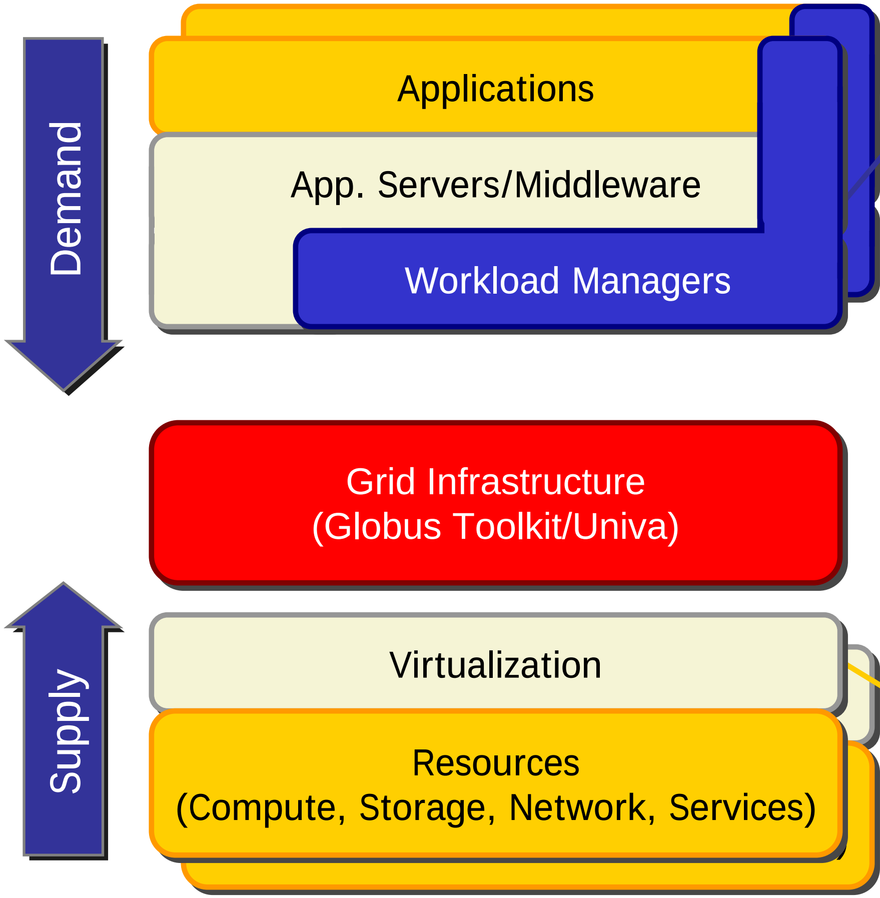
Grid Infrastructure: What is it?

- Provides…
  - Resource modeling
  - Reservation
  - Monitoring, notification
  - Provisioning, lifecycle management, decommissioning
  - Accounting, auditing
- …For multiple resource types…
  - Computers, storage, networks, software, …
- …Enabling
  - Increased flexibility, reduced cost (CapEx and OpEx)
  - Utility providers, software as service

Grid Applications ≠ Grid Infrastructure

	Grid Applications	Grid Infrastructure
Drivers	▫ Speed up applications ▫ Utilize “idle cycles”	▫ Manage the execution environment within policy constraints ▫ Deliver end-to-end QoS
Target	▫ Application workloads that are inherently parallelizable ▫ Example: Monte Carlo simulation	▫ Widely varying application workloads that do not require tight transactional coupling to databases ▫ Example: Data driven analytics/BI
Environment	▫ Single cluster, idle workstations, servers ▫ Single workload type ▫ Static provisioning	▫ Multi cluster, servers, workstations, storage, network devices, sensors ▫ Multiple workload types ▫ Dynamic (“On Demand”) provisioning
Footprint	▫ Vertically integrated execution stack ▫ Proprietary interfaces	▫ Horizontal, loosely coupled services ▫ Open, standards based interfaces

Grid Infrastructure in the Data Center "Stack"



Scheduling of varied (coarse or fine grained) tasks on behalf of multiple applications

Management of the resources...
Through discovery, reservation, enactment, monitoring...
To provide end-to-end QoS...
For multiple workload types

Isolation of resources to simplify provisioning and decommissioning

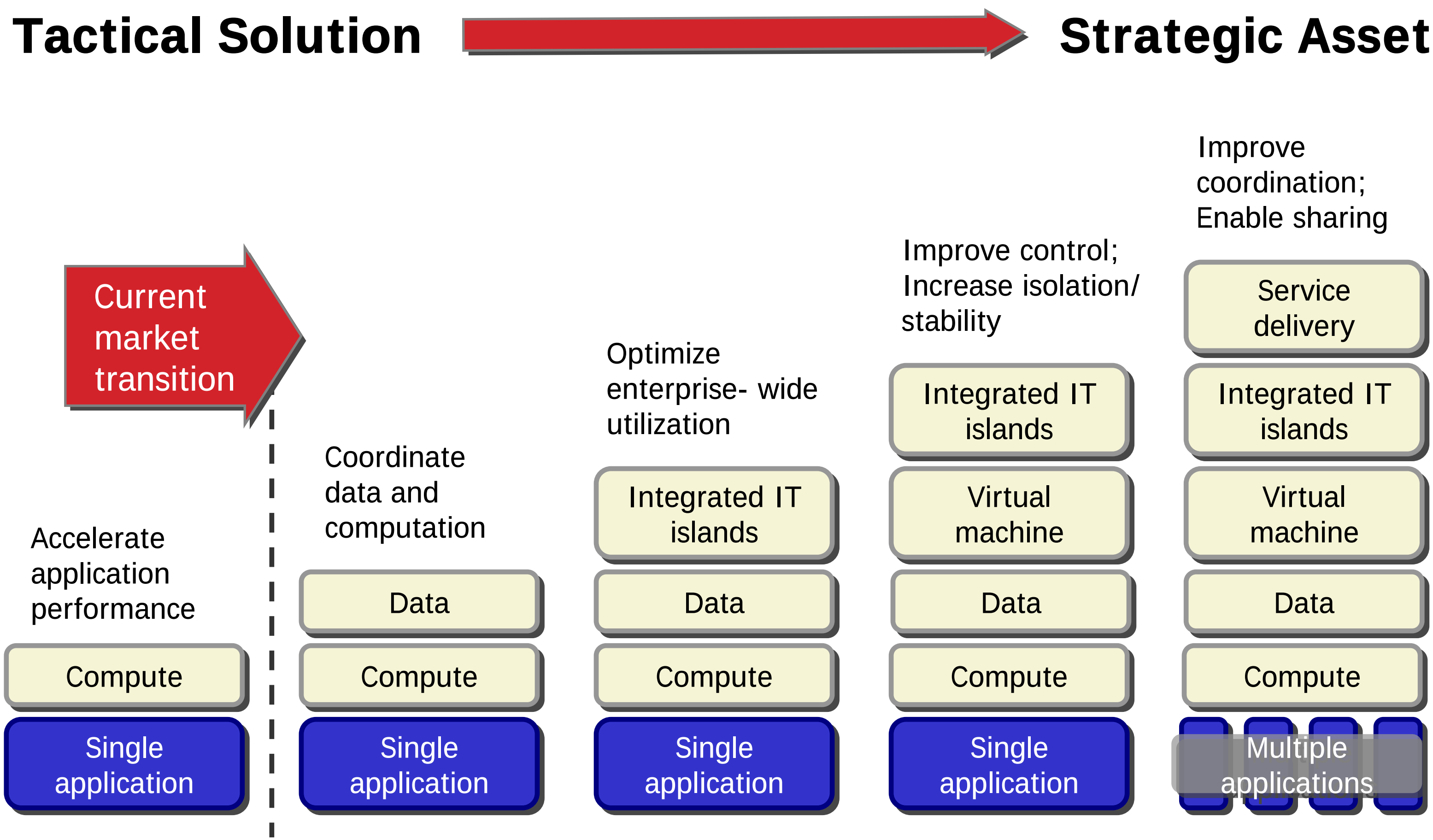
* Reference: The Different Faces of IT as Service, Ian Foster & Steven Tuecke, ACM Queue, July/August 2005

Grid Infrastructure Impact

Issue	Grid Infrastructure Impact
Hardware is either over- or under-utilized based on forecasting of previous application demands Non-static workload demands => impossible to cost-effectively utilize resources	Delivers ability to reserve and allocate resources for service delivery as needed Guarantees quality of service to all applications sharing a common resource pool
In large IT environments, typical server provisioning cycle is 3-4 *months*	Allows infrastructure to grow and shrink as application demands fluctuate
Unplanned downtime negatively impacts operations, ability to address service requests	Allows delivery to be continuously reallocated or "re-purposed", resulting in effective responsiveness

Reliable and responsive infrastructure reverses the "**erosion of trust**" between IT and the business

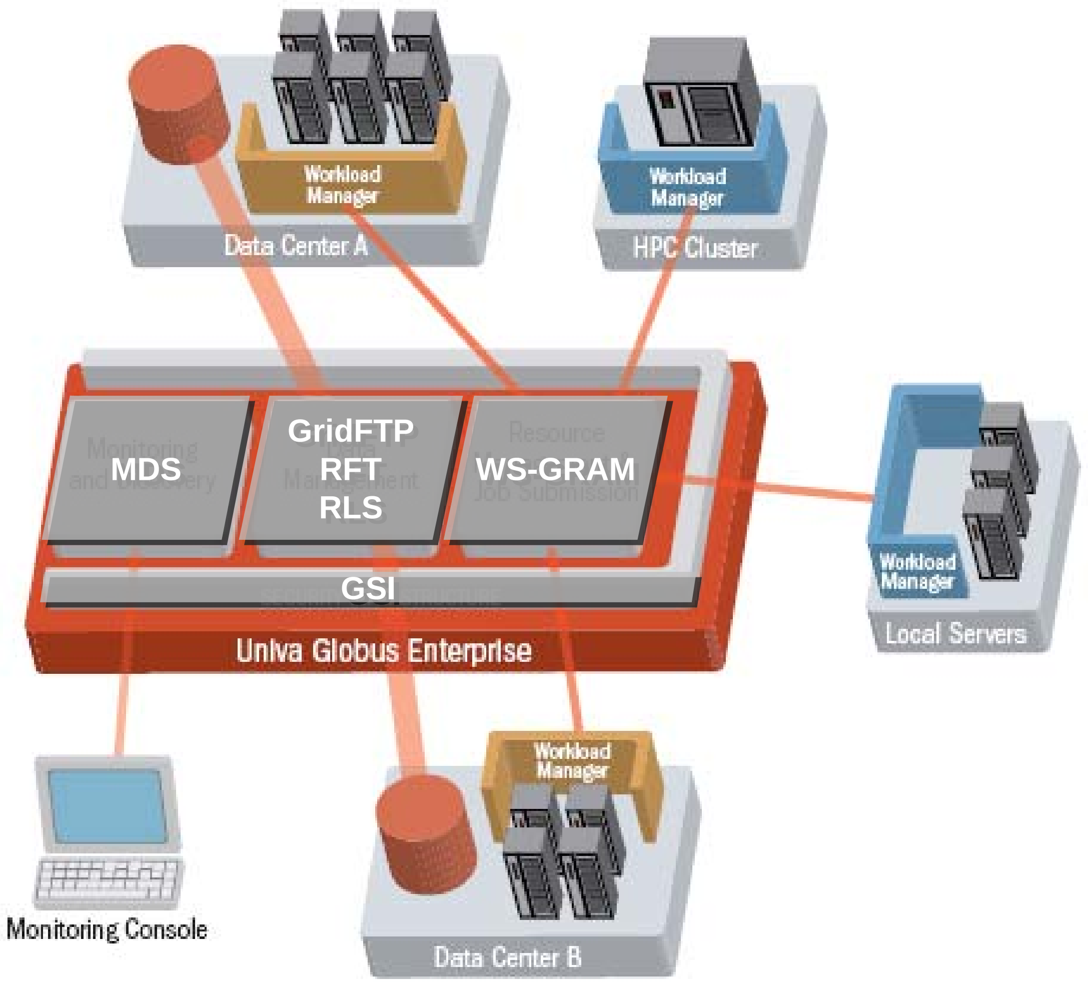
The Value Progression



Role of Open Source and Standards

- Companies wary of lock-in at the infrastructure level
- Pervasive nature of infrastructure creates constant change
- Management is beyond the scope of any single vendor
- Open source community allows vendor collaboration in a “non-threatening” forum
  - Drive open standards
  - Use open source as a proxy for emerging standards

Univa Globus Enterprise
Turning Open Source into Business Value



- Standard mechanisms for interfacing with multiple job management software
- High performance, reliable data transfers among Grid resources
- Scalable services for discovery and monitoring of resources within the Grid
- Enterprise-wide security functions such as single sign-on, authorization, and delegation of credentials

Univa is...

- Founded in 2004 by Steve Tuecke, Ian Foster and Carl Kesselman
  - Fathers of Grid & creators of the Globus® Toolkit
  - Based on 10 years of development and production use of Globus in over 1,000 global Grids
- Offering enterprise-class Grid infrastructure solutions
  - Grid infrastructure software
  - Professional services
  - Customer support
  - Channel partner community