Multicore Computing and the Cloud
Optimizing Systems with Virtualization

Michael Gschwind
IBM Systems and Technology Group
Major Forces Are Driving the Need For IT Transformation

Operational issues have IT at a breaking point

IT Infrastructure

The pace of technology innovations is accelerating

Costs & Service Delivery
- Rising costs of operations
- Pressure to reduce overall IT costs
- Explosion in volume of data and information
- Difficulty in deploying new applications & services

Business Resilience & Security
- Growing systems & applications availability needs
- Security of your assets & your clients’ information
- Landslide of compliance requirements

Energy Requirements
- Rising energy costs & rising energy demand
- Power & thermal issues inhibit operations
- Environmental compliance & governance mandates

Technology Advances
- Service orientation
- End-to-end service mgmt
- Comprehensive virtualization
- Converged networks
- Solid state storage
- IT appliances
- Storage de-duplication
- Many cores & threads per chip
- Low-cost high-BW fiber optics
- Petaflop supercomputers
- Cloud computing services
- Real-time data streams
- …
Integration In Power

Power 1
Circa 1990

Power Cluster
Mid 90’s

Power 4
Chip Multiprocessors (CMP)
Circa 2000

Cell
Power5 and better
Multi-core
Circa 2005-present

How to Value Integration in the Future
Ease of Use
Performance
Power / Price Reduction
Lower TCO

CPU
Chip
System
Data Center
Scaling Servers Up (and Down!) for Workload Needs

- Server scale-up driven by technology
  - Higher performance cores
  - Multicore technology

- Scaling up server size
  - improves performance
  - reduces management complexity
  - reduces TCO:
    improved performance/cost and reduced admin cost

- Not all workloads are scaling up with server size

- Partitioning hardware server ➔ multiple smaller virtual systems
Server Partitioning in Space and/or Time

- Partition a server across space and/or time

- Hardware partitioning
  - Better Performance Isolation
  - Limited allocation granularity

- Time Sharing
  - Fine grained allocation
  - Less Performance Isolation
Server virtualization approaches

Hardware Partitioning

- Server is subdivided into fractions each of which can run an OS
- Granularity coarse but improving
- Number of virtual machines is limited but may be sufficient
- Hardware fractions become dedicated to partitions
- Board-level partitioning may provide electrical isolation

Bare Metal Hypervisor

- Hypervisor software/firmware runs directly on server
+ Granularity may be very fine
+ Sum of virtual resources may far exceed physical resources
+ Dynamic timesharing of all physical resources
+ Virtual machines may have dedicated hardware resources

Hosted Hypervisor

- Hypervisor software runs on a host operating system
+ Granularity may be very fine
+ Sum of virtual resources may far exceed physical resources
+ Hypervisor can exploit OS services and device drivers

- Host OS is large single point of failure and overhead factor

- E.g.: HP nPars and vPars, Sun Domains and Logical Domains, BlueGene
- E.g.: PR/SM, z/VM, PowerVM, VMware ESX, MS Hyper-V, Xen
- E.g.: MS Virtual PC, Linux KVM, HP Integrity VM
Increasing Dynamic Capabilities of Partitioning

- From static partitioning at system init time…
  - Offer server cost benefit
  - Reduced administration complexity
  - Reduced server floor space

- … to flexible administration of partitioned systems at runtime
  - Increase operational efficiencies
  - Dynamic tuning to workload peaks
  - Reduce power consumption
  - Increase reliability and serviceability
Increasing value with dynamic partitioning capabilities

Partition Image Management

Partition Migration

Dynamic LPAR

LPAR

Improve server and application provisioning

Improve Reliability, Availability and Serviceability and Energy Efficiency

Improve operations with dynamic response to workload performance/resource needs

Improve Acquisition / TCO, static resource allocation

Increasing desirability
## Server Virtualization Hierarchy

<table>
<thead>
<tr>
<th>Virtual Resources</th>
<th>Implementation Methods</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Virtual runtimes</strong> (&lt;br&gt;application containers)&lt;br&gt;“Application Virtualization”&lt;br&gt; • Virtual operating systems&lt;br&gt; (application containers)</td>
<td>• Middleware provides JVM, J2EE, or CLR application containers&lt;br&gt; • Multiple middleware instances act as one</td>
<td>• WebSphere VE&lt;br&gt; • Microsoft and Sun JVMs&lt;br&gt; • BEA Liquid VM</td>
</tr>
<tr>
<td>• Virtual servers&lt;br&gt; (virtual machines / LPARs)&lt;br&gt; • Virtual I/O&lt;br&gt; • Virtual networks&lt;br&gt; • Virtual devices (CPUs, memory, I/O adapters, …)</td>
<td>• OS creates virtual OS environment per app.&lt;br&gt; • Each container has its own name space, files, root, …</td>
<td>• z/OS Address Spaces&lt;br&gt; • AIX 6.1 WPARs&lt;br&gt; • Solaris Containers&lt;br&gt; • HP-UX Syst. Res. Part.&lt;br&gt; • MS SoftGrid</td>
</tr>
<tr>
<td>• Hypervisors&lt;br&gt; • HW partitioning&lt;br&gt; • Virtual I/O Servers&lt;br&gt; • Self-virtualizing I/O adapters&lt;br&gt; • In-memory VLANs</td>
<td>• System z PR/SM and z/VM&lt;br&gt; • PowerVM Hypervisor&lt;br&gt; • HP vPars, Integrity VM&lt;br&gt; • Sun Logical Domains&lt;br&gt; • VMware, MS Hyper-V, Xen, KVM, Virtual Iron, …</td>
<td></td>
</tr>
</tbody>
</table>
Server Virtualization Hierarchy

Virtual Resources | Benefits | Issues
--- | --- | ---
- **Virtual runtimes** (application containers)  
  "Application Virtualization"  
- **Virtual operating systems**  
  (application containers)  
- **Virtual servers** (virtual machines / LPARs)  
  **Virtual I/O**  
  **Virtual networks**  
  **Virtual devices** (CPUs, memory, I/O adapters, …)  
- **Increased hardware utilization ➔ less HW cost, power used, and floor space**  
  **Agility / flexibility**  
  **Fine granularity**  
  **High efficiency**  
- **Fewer OS kernels**  
  **Extremely fine granularity**  
  **Improved efficiency via single shared OS**  
- **Virtual runtimes can be independent of OS and hardware**

- **Container management**  
- **All apps must like same middleware release**  
- **Isolation at middleware level only**  
- **Container management**  
- **All apps must like same OS release**  
- **OS service affects all apps**  
- **Isolation at OS level only**  
- **Avail. & security depends upon hypervisor design and integration**  
- **Doesn’t reduce # of OS images to be managed**  
- **Introduces licensing and usage accounting issues**

System Stack

- **Applications**  
  - Most special purpose  
  - Most lightweight  
  - Most HW independent

- **Middleware**

- **Operating Systems**
  - Most general purpose  
  - Most heavyweight  
  - Most HW sensitive

- **Hardware**

- **Hardware**
  - Most special purpose  
  - Most lightweight  
  - Most HW independent
Hardware facilities: challenges and opportunities

- Three differentiated privilege levels desirable
  - application, OS, hypervisor

- Virtual machine support in HW
  - Native hardware support to virtualize a guest in hardware
    - System z operating systems use fully virtualized hardware
    - Paravirtualization provides a collaborative framework between HV and OS
      - OS sends requests to HV via HCALLs

- Virtualize instruction sets, reduce friction in deployment
  - IBM work on architecture virtualization (DAISY, BOA) in 90’s
    - Supporting multiple architectures on one common core
    - Flexibility to respond to demand for different architectures
  - Today, PowerVM has Lx86 option to virtualize x86 server on Power HW
Multicore optimization challenges and opportunities

- QoS guarantees for performance isolation between co-resident partitions
  - Memory bandwidth and on-chip caches are critical resources

- Resource Allocation and Resource Metering
  - Better control over allocation of resources to partitions and billing for their use

- Exploit in-server locality for improved network performance
  - “Virtual network” as memory-to-memory copy
Virtualization beyond the Processor Domain

- **Virtualization and partitioning interacts with I/O and networking**
  - Native traditional I/O into physical memory risks breach of partition isolation
  - Partition migration requires that I/O and networks follow the moving partition

- **Virtualization of networks and I/O resources will increase**
  - Virtual networks and virtual I/O adapters are becoming virtual resource objects that can be provisioned, moved while active, … like their server and storage counterparts

- **Networks and I/O adapters will provide their own self-virtualization capabilities**
  - I/O adapters are implementing virtualization in hardware
    - Starting with System Z (80’s) and Power (2007)
    - Trickling down into commodity standards with NPIV, SR-IOV, …
    - Ethernet is evolving to support VLANs and virtual Fibre Channel networks (VSANs)
Cloud Computing is a new way to view systems

- Transition from hardware pieces to execution object (corresponding to a virtual system or partition)
  - Not pieces of physical hardware, but objects with the capability to execute a task

- The building blocks of Cloud Computing are partitions, virtual storage, virtual I/O, virtual networks...

- Cloud Computing is a way to better manage these objects and allow better scale-up in increasingly large environments
  - Allocate
  - Provision
  - Configure
  - Initialize
  - Optimize usage
Stages of IT Infrastructure Transformation

Virtualization has major benefits but introduces more complexity

Consolidation

Objects and Pools

Virtual Resource Objects (Servers, Storage, Networks)

Multi-System Virtualization

Resource Pools

• Reduced scale-out complexity
• Integrated autonomic mgmt
• Dynamic energy optimization
• Greater resilience / security

Integration for Simplification

• Better SW investment protection
• Simplified HA solutions
• Improved resource optimization
• Ready-to-run packaged software

Continuing Advances

Continuing Advances (I/O, network, resilience, performance, …)
Ensemble Components

An ensemble generally consists of the following components:

- A pool of compatible system nodes (e.g., N standard servers; typically homogeneous)
- Virtual resource mobility within an ensemble and with compatible ensembles
- The networks which interconnect the ensemble nodes (may be local / optimized)
- Resource virtualizers (hypervisors, I/O virtualizers, storage virtualizers, …)
- An ensemble manager function that provides platform management for the ensemble virtual and physical resources
- Tools for planning, ensemble creation, P2V migration, image mgmt. & composition, …
- Ensemble-local automated optimization software of performance, availability, energy usage, security, … with intelligent defaults
- Multi-system services (locking, caching, message queuing, …) may be provided
A Shift from Managing Virtualization to using Virtualization to Manage...

- **Today’s Hardware Centric Approach**
  - Carve up machines – partitioning and sharing of the hardware
  - Low level knobs and dials that are different on every platform
  - Bottoms-up build-out of the virtualized infrastructure

- **Tomorrow’s Partition Image Centric Approach and Value**
  - A single and consistent solution for all systems
  - Radically improved time to value for new solutions
  - Provide repeatable accuracy and consistency via automation
  - Drive higher utilization and efficiencies of systems
  - Significantly reducing energy costs of systems
  - Provide improved business application workload availability
  - Support an ecosystem around delivery of SaaS in the cloud
Supporting an Ecosystem around Standardized Virtual Appliances

- **A Partition Image is a Virtual Appliance**
  - Simplifies delivery and deployment of complex software systems
  - Provides pre-installed, pre-configured and tested software stack
    - Operating Systems, Middleware and Application Software
  - The entire software solution is managed (deployed, updated, etc.) as a unit
    - Removes the need to deal with problematic dependency management
The Detailed view of a Virtual Appliance

- A Virtual Appliance is a self-contained package containing….
  - meta-data describing the required server resources needed to run the image
    - Number of CPUs (dedicated vs. shared)
    - Memory requirements
    - IO and network requirements
  - meta-data describing goals and constraints for the image
    - Performance and availability goals
    - Placement constraints (i.e. security isolation)
  - meta-data describing configuration variables
    - OS Configuration parameter – IP Address, etc.
    - Application Configuration parameters
  - One or more disk images containing OS, middleware and other application software
Managing Deployment of a Partition Image (Virtual Appliance)

- **Deploy a Virtual Appliance**
  - The Virtual Appliance meta-data is used to create VM container, allocating the required platform resources.
  - The bootable disk image is copied and/or made accessible to the selected host system.
  - The virtual machine is started from the bootable disk image and customized as part of its initial boot.

- Improved time-to-value…
- Fewer tools and fewer tasks…
- Workload (business) context…
Deploying A Virtual Server Collection (Composition)

- A Virtual Appliance representing a multi-tiered applications workload is selected for deployment.
- Understanding of the composition is maintained and used in the management of the running workload.
- Intelligent placement of the virtual machines within the pool of systems provides simplification and optimizations.
Automating Workload Availability for Unplanned Downtime

- Move virtual servers away from a failing host system.
  - Use of HW Predicted Failure Analysis to drive ‘live’ virtual machine mobility.

- Restart virtual servers when a host system fails.
  - Restarting a virtual server (possibly from checkpoint) on another server is a form of ‘static’ migration.
Non-Disruptive System Updates

- Firmware updates are applied to the pool as a single system with no disruption to the running workload.

- The System Pool manager orchestrates the movement of the deployed virtual machines as the updates are applied to each of the required systems.

- Placement Services ensures the optimal placement for each virtual machine as they are relocated.
Automated Energy Optimizations

- Consolidate Virtual Servers on a fewer number of host systems.
  - Move using ‘live’ virtual machine mobility (relocation).

- Power Off / Suspend host systems that are currently not required.
Cloud Computing – a user’s perspective

Cloud Computing is a style of computing in which applications, data, platforms, and resources are provided as services to users over the Web.

- The services provided may be available globally, always on, low in cost, “on demand”, massively scalable, “pay as you grow”, …
- Consumers of the services need only care about what the service does for them, not how it is implemented.
Types of Clouds and Cloud Services

Public Clouds
(provider - Internet)

Private Clouds
(data center - Intranet)

Hybrid Clouds
(public and private)

‘Services’ as a service

Applications, processes and information as a service

Software platforms as a service
(optimized middleware – application servers, database servers, portal servers, etc.)

Infrastructure as a service
(virtualized servers, storage, networks)
IT Clouds Will Be Aggregations Of Smaller Clouds

IT Cloud:
- Heterogeneous with many traditional networked systems (not shown)
- Immense scale (may even span datacenters)
- Service Management
- Network service (data center fabric)
- Security service (federated identity)
- Service Registry

Middleware Ensembles

JAVA Application
Server service

Database service

File system service

Storage backup, archive ... service

Storage Ensembles

Network/Security Ensembles

Consolidation service

Virtual Client service

ERP Application service

DMZ appliance service
Research Computing Cloud (RC²) - Overview

- A ‘living lab’ to advance research strategies
- Provides self service ‘on demand’ delivery solution for research computing resources
- Integrates existing assets and products using SOA
- Leverages off-the-shelf (Tivoli) products
- Zero touch support for the full life cycle of service delivery
  - Order creation
  - Approval process
  - E-mail notification
  - Automated provisioning
  - Monitoring
Summary

- Virtualization technology is key for cloud computing
- Cloud Computing transcends static view of a system as a piece of hardware and views systems as dynamic entities
- The increased flexibility opens new ways to think of systems
  - Break link between system and HW rack
  - Can move system from one rack to another
    - Reliability, availability, serviceability, energy optimization, …
    - Can pre-configure a system without rack
- Focus on how systems are used, not how they are built
- Offer increased efficiency by making systems more dynamic
- Increase the utility of systems to the users
- Make IT more consumable