IBM and Next Generation Linux

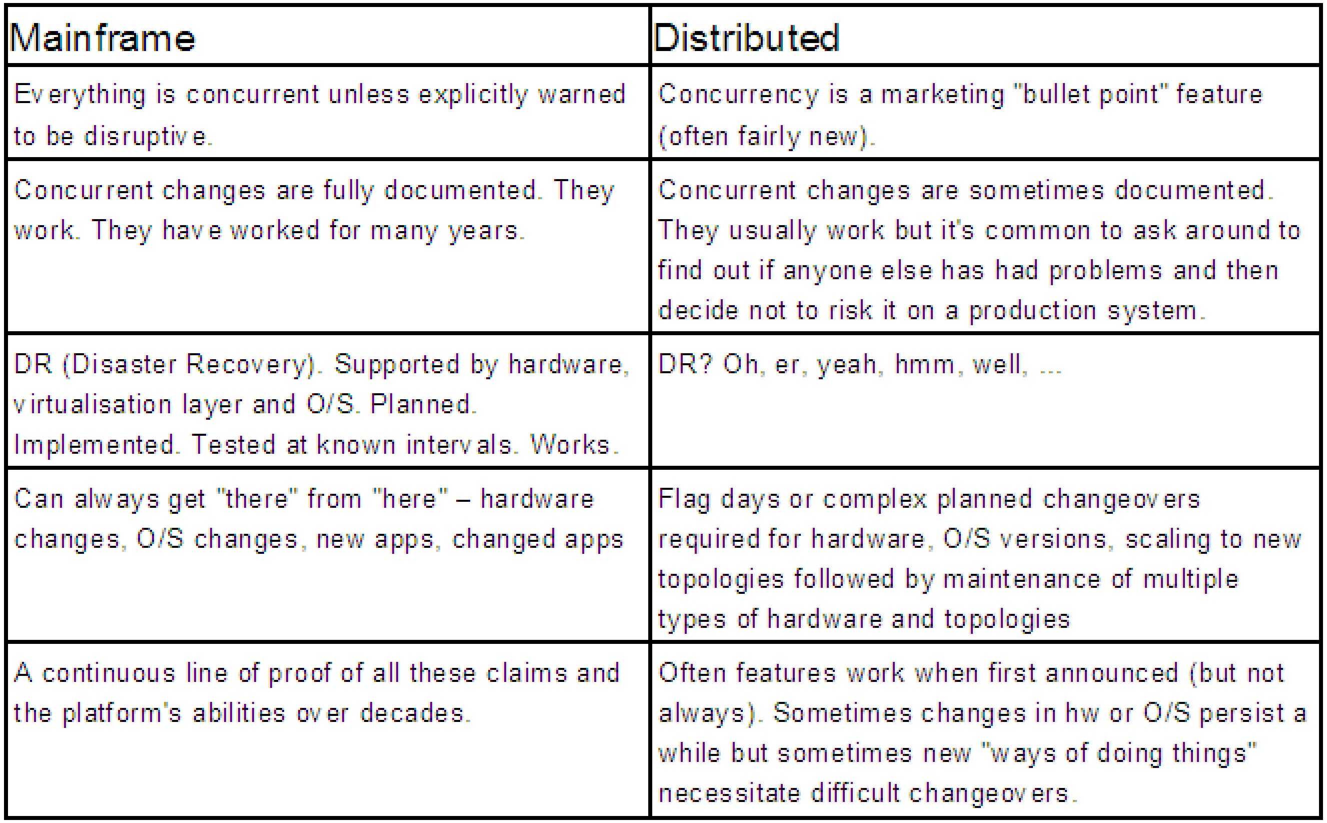
Linux on System z

Europe System z software & hardware business

Agenda

- System z hardware
- System z virtualisation
- Linux on System z
- Questions

Points of View



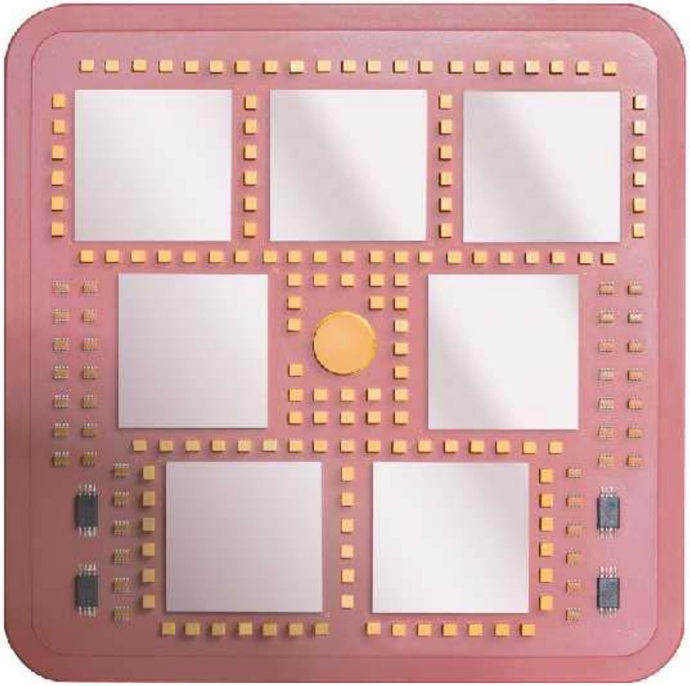
Mainframe	Distributed
Everything is concurrent unless explicitly warned to be disruptive.	Concurrency is a marketing "bullet point" feature (often fairly new).
Concurrent changes are fully documented. They work. They have worked for many years.	Concurrent changes are sometimes documented. They usually work but it's common to ask around to find out if anyone else has had problems and then decide not to risk it on a production system.
DR (Disaster Recovery). Supported by hardware, virtualisation layer and O/S. Planned. Implemented. Tested at known intervals. Works.	DR? Oh, er, yeah, hmm, well, ...
Can always get "there" from "here" – hardware changes, O/S changes, new apps, changed apps	Flag days or complex planned changeovers required for hardware, O/S versions, scaling to new topologies followed by maintenance of multiple types of hardware and topologies
A continuous line of proof of all these claims and the platform's abilities over decades.	Often features work when first announced (but not always). Sometimes changes in hw or O/S persist a while but sometimes new "ways of doing things" necessitate difficult changeovers.

System z10 Overview

- Machine Type
  - 2097
- 5 Models
  - E12, E26, E40, E56 and E64
- Processor Units (PUs)
  - 17 (17 and 20 for Model E64) PU cores per book
  - Up to 11 SAPs per system, standard
  - 2 spares designated per system
  - Dependent on the H/W model - up to 12, 26, 40, 56 or 64 PU cores available for characterization
  - Central Processors (CPs), Integrated Facility for Linux (IFLs), Internal Coupling Facility (ICFs), System z10 Application Assist Processors (zAAPs), System z10 Integrated Information Processor (zIIP), optional - additional System Assist Processors (SAPs)
- Memory
  - System Minimum of 16 GB
  - Up to 384 GB per book
  - Up to 1.5 TB for System and up to 1 TB per LPAR
  - Fixed HSA, standard
  - 16/32/48/64 GB increments
- I/O
  - Up to 48 I/O Interconnects per System @ 6 GBps each
  - Up to 4 Logical Channel Subsystems (LCSSs)
- ETR Feature, standard

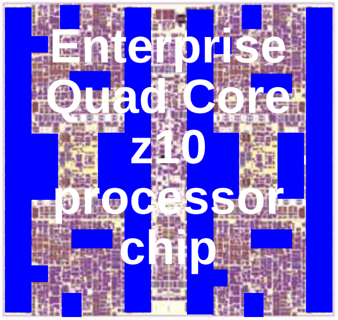
z10 EC Multi-Chip Module (MCM)

- **96mm x 96mm MCM**
  - 103 Glass Ceramic layers
  - 7 chip sites
  - 7356 LGA connections
  - 17 and 20 way MCMs

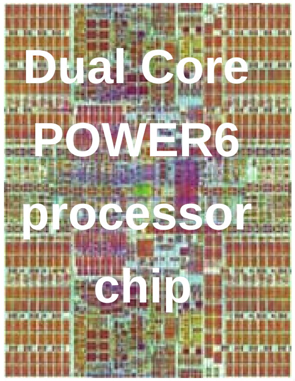


- **CMOS 11s chip Technology**
  - PU, SC, S chips, 65 nm
  - 5 PU chips/MCM – Each up to 4 cores
    - One memory control (MC) per PU chip
    - 21.97 mm x 21.17 mm
    - 994 million transistors/PU chip
    - L1 cache/PU core
      - 64 KB I-cache
      - 128 KB D-cache
    - L1.5 cache/PU core
      - 3 MB
    - 4.4 GHz
    - 0.23 ns Cycle Time
    - 6 km of wire
  - 2 Storage Control (SC) chip
    - 21.11 mm x 21.71 mm
    - 1.6 billion transistors/chip
    - L2 Cache 24 MB per SC chip (48 MB/Book)
    - L2 access to/from other MCMs
    - 3 km of wire
  - 4 SEEPROM (S) chips
    - 2 x active and 2 x redundant
    - Product data for MCM, chips and other engineering information
  - Clock Functions – distributed across PU and SC chips
    - Master Time-of-Day (TOD) and 9037 (ETR) functions are on the SC

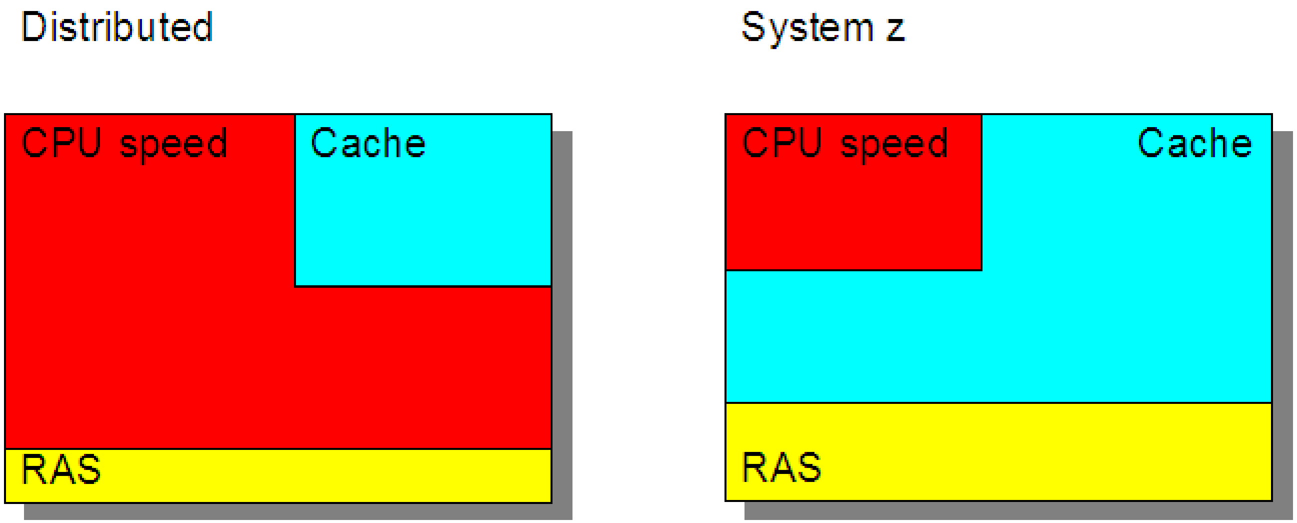
z10 EC Chip Relationship to POWER6™



- **New Enterprise Quad Core z10 EC processor chip**
- **Siblings, not identical twins**
- **Share lots of DNA**
  - IBM 65nm Silicon-On-Insulator (SOI) technology
  - Design building blocks:
    - Latches, SRAMs, regfiles, dataflow elements
  - Large portions of Fixed Point Unit (FXU), Binary Floating-point Unit. (BFU), Hardware Decimal Floating-point Unit (HDFU), Memory Controller (MC), I/O Bus Controller (GX)
  - Core pipeline design style
    - High-frequency, low-latency, mostly-in-order
  - Many System z and System p designers and engineers working together
- **Different personalities**
  - Very different Instruction Set Architectures (ISAs)
    - very different cores
  - Cache hierarchy and coherency model
  - SMP topology and protocol
  - Chip organisation
  - IBM z10 EC Chip optimised for Enterprise Data Serving Hub

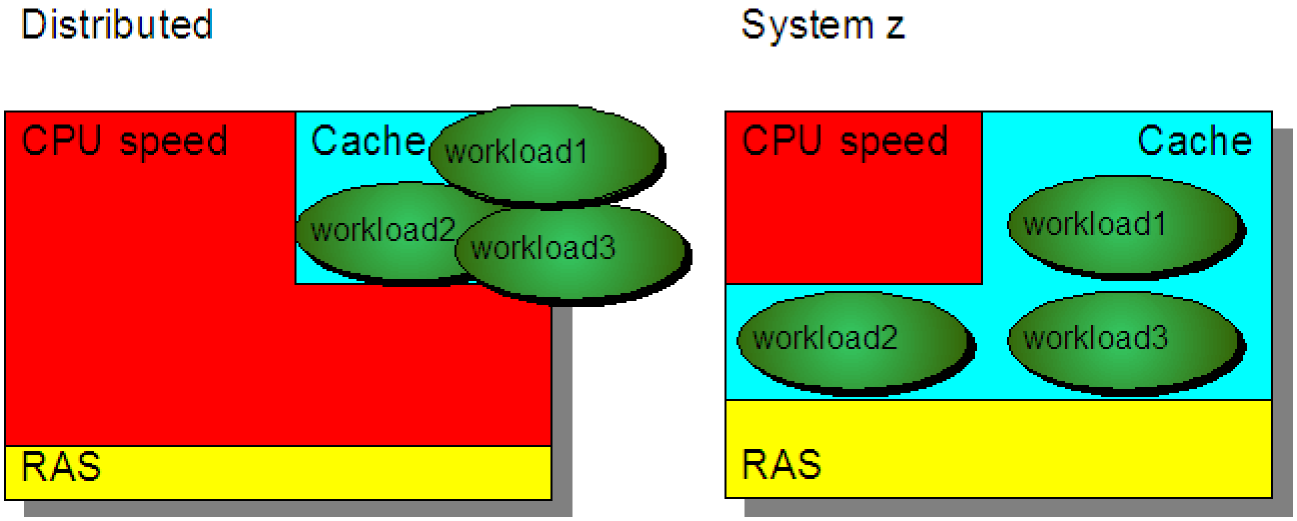


Processor design points: chip real estate



Processor design points: chip real estate

Running mixed workloads



- Working sets may be too large to fit in cache
- "Fast" processor speed is not realised due to cache misses
- RAS failures have more severe consequences

- System z cache is able to contain more working sets
- Processor speed is optimised by increased cache usage
- Additional RAS function is beneficial for mixed workloads

I/O Connectivity

- FICON adapters are used for connectivity to disk and tape
  - Up to 336 FICON channels can be installed in a z10 EC
  - Latest generation is FICON Express4, channel is full duplex at 4 Gbs
    - Up to 350 MByte/s throughput (large I/Os); up to 13000 I/O per second (4KB I/Os)
  - Each channel can be put in either
    - FICON mode – traditional mainframe protocol which includes built-in transparent multipathing and measurement support
    - FCP ("Fibre Channel Protocol") mode to provide SCSI SAN device access to Linux on System z, z/VM and z/VSE
- A range of OSA-Express2 features support ethernet connectivity
  - 10 Gbps over fibre (up to 10km unrepeated)
  - 1 Gbps over fibre (up to 10km unrepeated) or 1000BASE-T (copper)
  - Some other network connectivity options are available too
  - Up to 48 ports per system
- Hipersockets
  - Allows partition-to-partition TCP/IP networking within the System z CEC
  - Extremely high performance and low latency (memory speed)
  - Extremely strong isolation assurance (EAL5, see Common Criteria section)

Cryptography

- CPACF: CP Assist for Cryptographic function
  - Symmetric, clear-key crypto algorithms are built into each CPU chip
    - DES, T-DES, AES-128, -192, -256, SHA-1, SHA-2 (224, 256, 384, 512), PRNG
- Other crypto can be offloaded to Crypto Express2 card features
  - Up to 8 features, each with 2 PCI-X cryptographic adapters
  - Each PCI-X crypto adapter can be configured as accelerator or co-processor
  - Co-processor
    - Public and secret key crypto
    - Tamper-proof, secure key, FIPS 140-2 Level 4
    - Algorithms such as T-DES, RSA, PIN
    - Supports Master Key or TKE (Trusted Key Entry) workstation
  - Accelerator
    - Clear-key, RSA operations
    - 6000 SSL handshakes per second (with one Crypto Express2 feature)

RAS characteristics of System z- Reliability, Availability and Serviceability

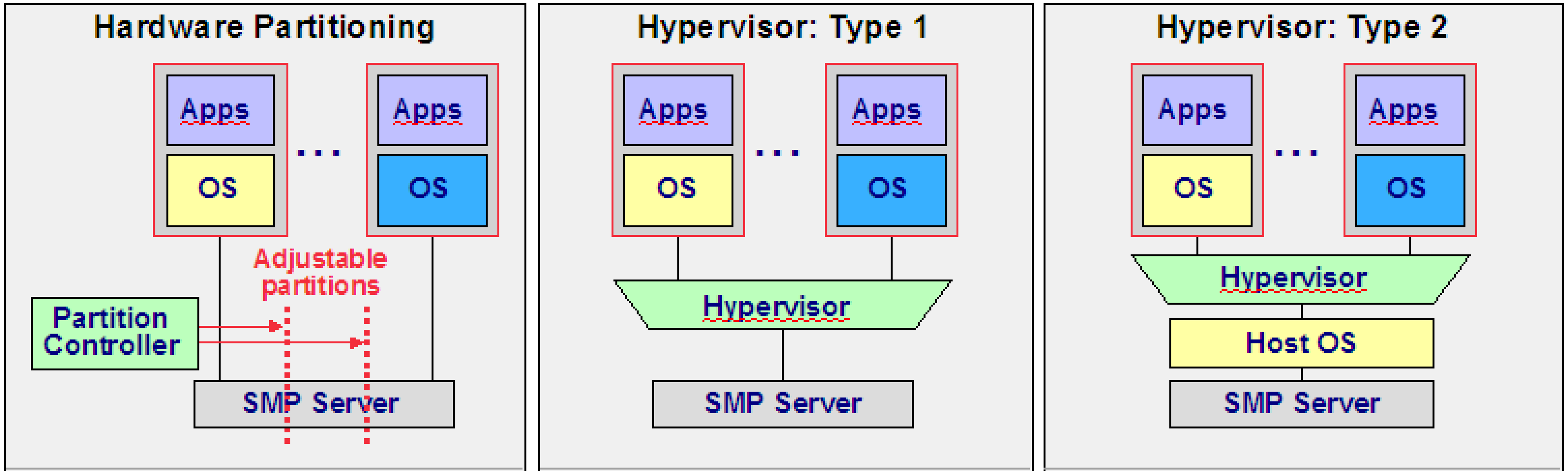
- **Each PU has dual instruction/execution engines: the two engines execute each instruction in lockstep;** output is compared at each checkpoint; on mismatch, PU retries from checkpoint; on continued failure, dynamic sparing occurs to any unused PU ("self-initiated brain transplant")
- **L1 cache parity checked and store-through to L2; supports cache-line delete (compartment and one-line),** cache-line sparing ("fuse" relocation)
- **L2 cache supports cache-line relocation; specially designed ECC (8 check bits for each 64-bit word),** Directory address field: (25,19) ECC; Directory ownership field: (11,5) ECC; Single error correction; double error detection; L2 cache can delete any combination of lines
- **Main memory: special ECC; (140,128) ECC code: Corrects single bit failures, single symbol failures (detects** any arbitrarily "flaky" chip); chips reserved for sparing
- **Address protection and failure isolation: special (144,132) ECC code uses 4 extra bits for memory address** protection (against erroneous fetch) and failure isolation (special bit patterns for cache/non-memory failure, memory error, memory store interface error)
- **DRAM with many defects are spared; DRAMs with high failure rates are spared; spares can be spared;** constant memory scrubbing; error counts accumulated; key protection (3 copies of every key (plus parity))
- **Concurrent replacement/repair - all cards can be replaced concurrently**
- **First error data capture (FEDC); Error logs stored in HMS or SE and queued to RETAIN**
- **Channel subsystem (CSS): STI and channel independence; Multipathing; Measurement; Administration; One** hardware instruction issues I/O request: other processors handle all the details
- **Common I/O Platform (CIOP) used for FICON-Express OSA-Express: PCI to STI adaption bridge; Dual cross-**checked PowerPC microprocessors and bridges; two independent data-mover queue (DMQ) engines

FAR TOO MANY TO MENTION IN 1 SLIDE

System z Security: Common Criteria

- Common Criteria
  - is an accepted standard for evaluating the security of a computing system
  - is based on a set of functional and assurance requirements
- A higher EAL rating assures higher security *for the tested security profile*
  - Common security profiles are
    - CAPP: Controlled Access Protection Profile
      - provides Discretionary Access Control ("DAC") – the "usual" profile
    - LSPP: Labelled Security Protection Profile
      - Provides "multi-level security" with Mandatory Access Control (MAC)
- The security requirements in Common Criteria have gained support as “best practices”
- PR/SM (the code that implements LPAR) is assured at level EAL5
- z/OS is assured at level EAL4+ for LSPP and CAPP
- z/VM 5.3 has completed evaluation at level EAL4+ for CAPP and LSPP
- Linux distributions SLES10 and RHEL4 are assured at level EAL4+ for CAPP

Server Virtualisation Approaches



Physical partitioning
Sun Domains, HP nPartitions

Logical partitioning
System p LPAR, HP vPartitions

Hypervisor software/firmware runs directly on server
System z PR/SM and z/VM
POWER Hypervisor
HP Integrity VM
VMware Infrastructure (ESX)

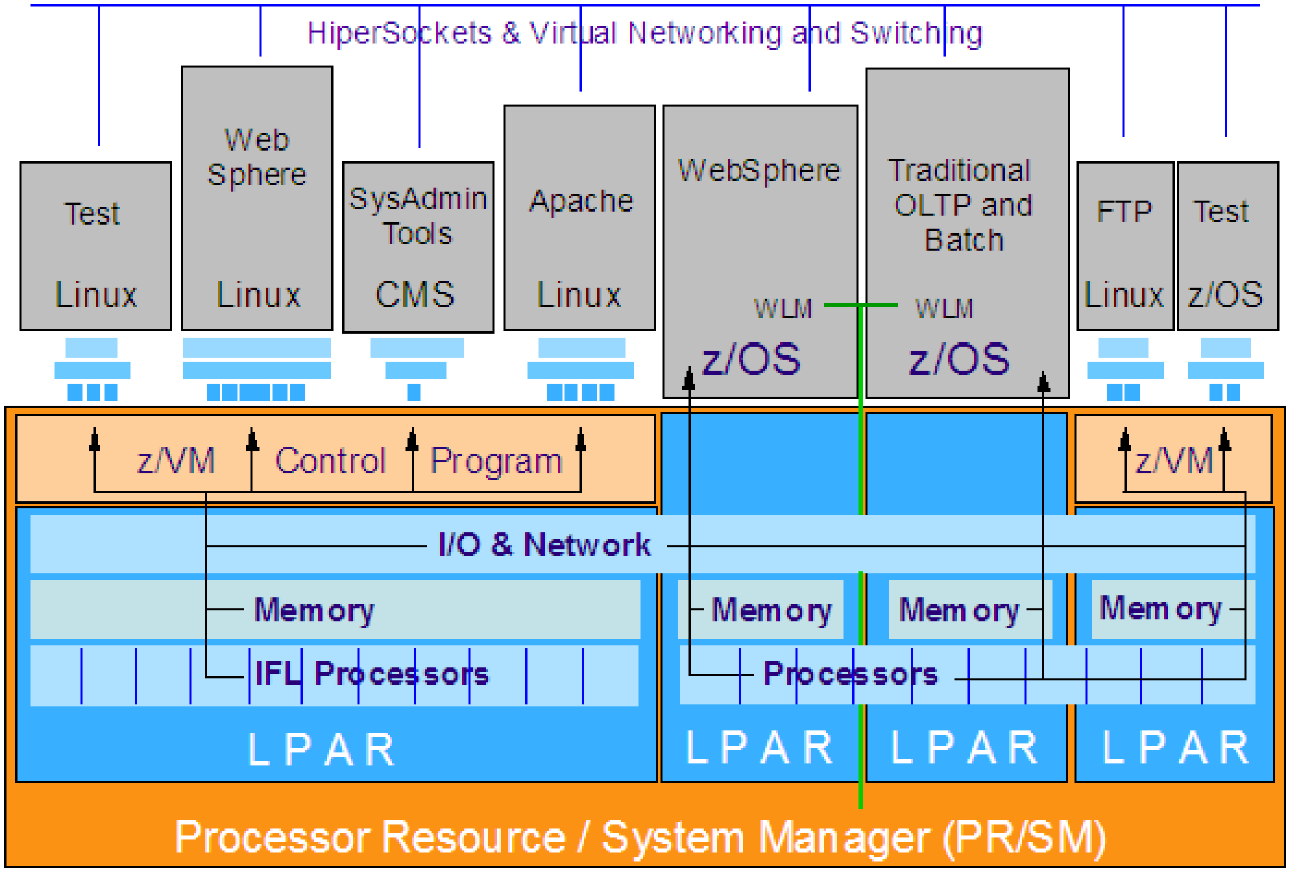
Hypervisor software runs on a host operating system
VMware Server (GSX)
Xen Hypervisor
Microsoft Virtual Server

- Hardware partitioning subdivides a server into fractions, each of which can run an OS
- Hypervisors use a thin layer of code to achieve fine-grained, dynamic resource sharing
- Type 1 hypervisors with high efficiency and availability will become dominant for servers
- Type 2 hypervisors will be mainly for clients where host OS integration is desirable

LPAR and z/VM

- Mainframe Logical Partitioning (LPAR), introduced in 1988, has provided years of business-critical, high-performance server partitioning for the world's largest corporations
  - Share processor, memory, I/O, and network among multiple operating environments
  - Isolates workloads with Common Criteria EAL5 assurance
- z/VM, commercially available since 1972, has supported mixed workloads that require minimal hypervisor overhead, massive scalability, and exceptional levels of availability
- Both LPAR and z/VM employ hardware and firmware innovations developed over the years that make virtualisation a fundamental part of IBM System z architecture

System z virtualisation / partitioning



- System z provides two levels of partitioning
- PR/SM enables scalable virtual server hosting for LPAR environments
- z/VM provides hypervisor function for highly scalable virtualisation

Scalability Limits

z/VM V5.3 on a single System z10 supports...

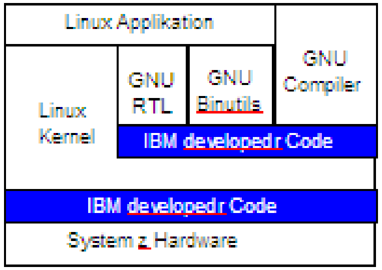
- up to thousands of guests (subject to physical resources, of course)
- up to 64 CPUs per virtual machine
- up to 32 (physical) CPUs per z/VM LPAR
- virtual machines with up to 256 GB of memory
- up to 256GB (physical) memory per z/VM LPAR
- virtual machines with I/O-intensive workloads
  - mainframe I/O bandwidth in a z/VM environment is very large and hardware assisted
- a mix of 31-bit and 64-bit guest images (and has done since 2001)
- non-disruptive addition and configuration of system resources such as CPUs, I/O devices, and network adapters to capable guests (including Linux, z/OS and z/VM)
  - and the hardware and z/VM supports non-disruptive addition and configuration of real CPU, I/O and network resources too

When Do You Need More Than "Good Enough"?

- ***Making the Case for Mainframe Virtualisation***
- When workload growth and decline is difficult to predict (be it production, development, or test and assurance systems)
- When business results suffer as a result of IT resources not matching customer demand
- When your IT staff wants to optimise their productivity for deploying and managing virtual servers
- When innovation is stifled because your staff cannot experiment or develop new solutions using existing resources
- When speed to market affects your business results
- When your server applications need fast and flexible access to z/OS data and applications
- When business continuance is a high priority
- When you want more control over your environmental expenses (e.g., floor space, cooling)

What is Linux on System z?

- **Standard Linux with ASCII** environment
- **Uses System z hardware, including IEEE floating point, HiperSockets,** traditional (ECKD) and FCP (Open Fibre Channel, SCSI) disks, CPACF cryptographic CPU instructions, …
- **Runs native, within an LPAR, or under z/VM**



Linux Applikation

GNU Compiler

Linux Kernel

GNU RTL

GNU Binutils

IBM developedr Code

IBM developedr Code

System z Hardware

Delta as low as 5%

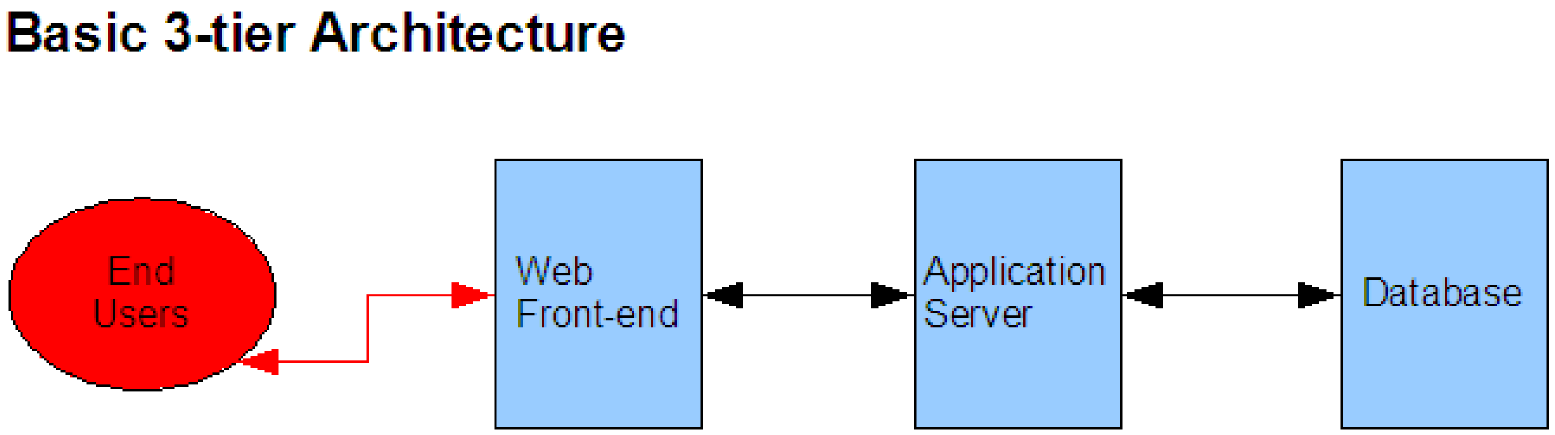
- **Design Principles**
  - Not a unique version of Linux (Linux is Linux is Linux)
  - No change to Look & Feel of Linux on System z
  - The IBM commitment to z/OS, z/TPF, and z/VSE is not affected

Integrated Facility for Linux (IFL)

- Engines (CPUs) characterised to run only Linux workloads
  - Supports z/VM and Linux on System z
  - Priced much cheaper than engines for z/OS use
  - IFLs on “sub-uni” systems run at “full speed”
    - z800, z890, z9 EC, z9 BC, z10
- Traditional mainframe software charges unaffected
  - IBM mainframe software
  - Independent Software Vendor products
- Linux and z/VM charged only against the IFL’s
  - Includes Linux and z/VM themselves and Linux software
  - Some IBM software also permits subcapacity pricing
    - pay for only the engines of guests that run the software

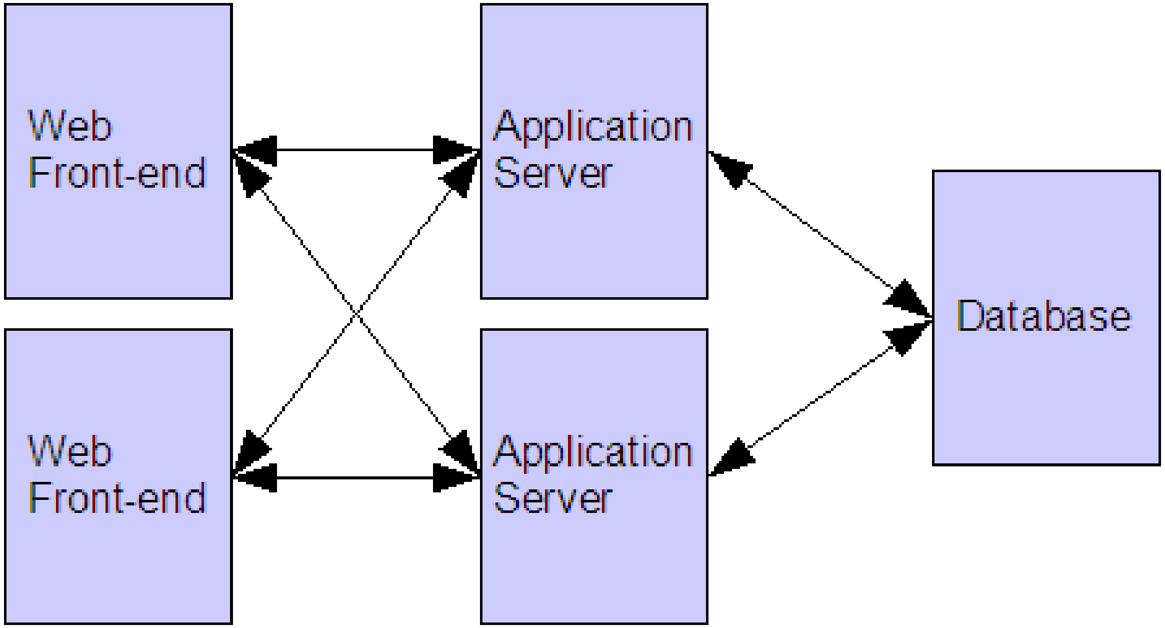
Where do all those servers come from?

Typical Enterprise Application Architecture



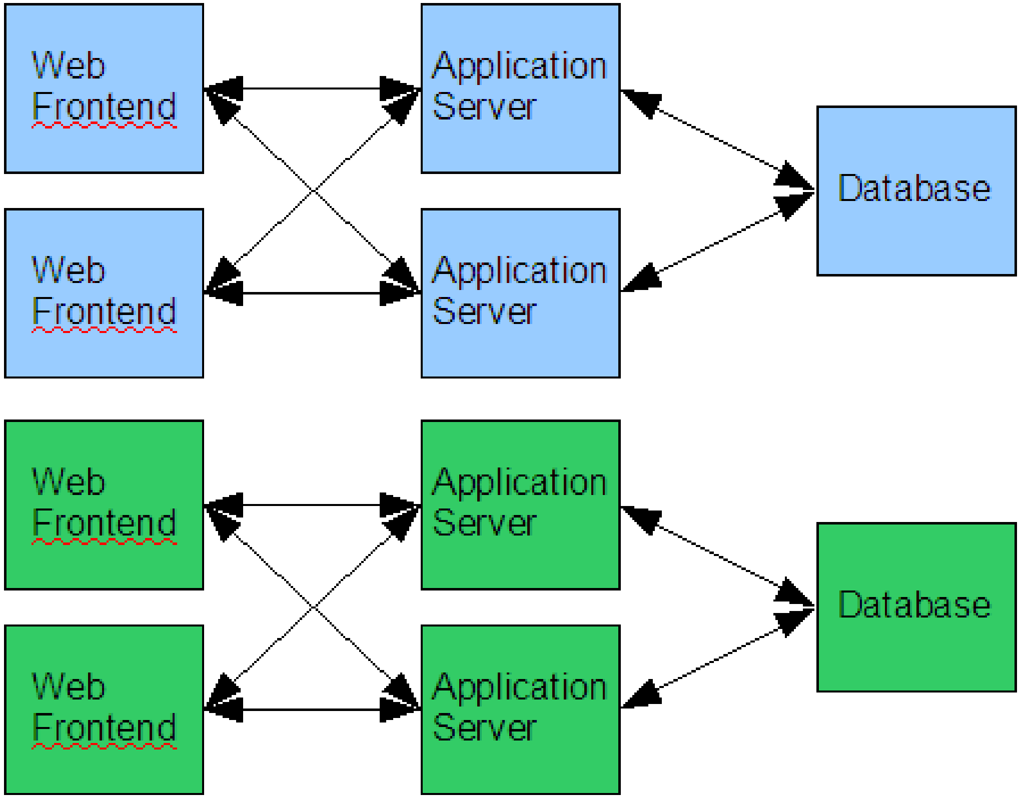
Typical Enterprise Application Architecture

Production deployment



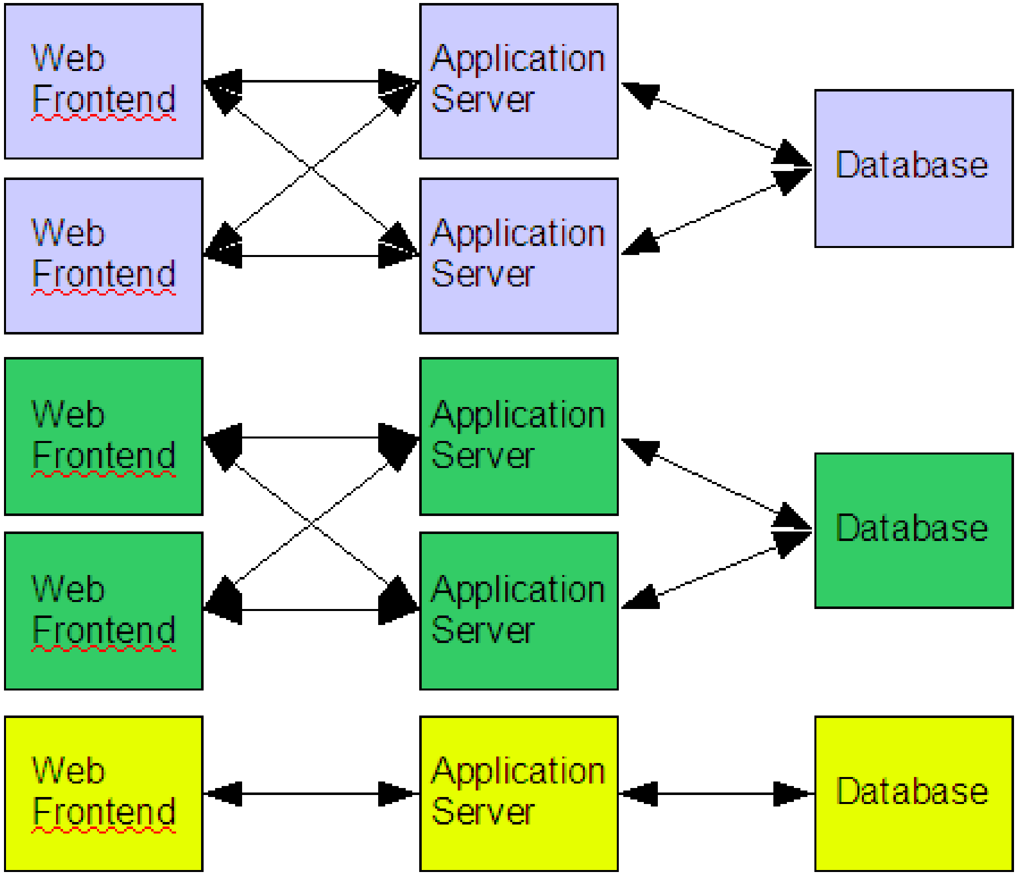
Typical Enterprise Application Architecture

Production plus System Test/QA



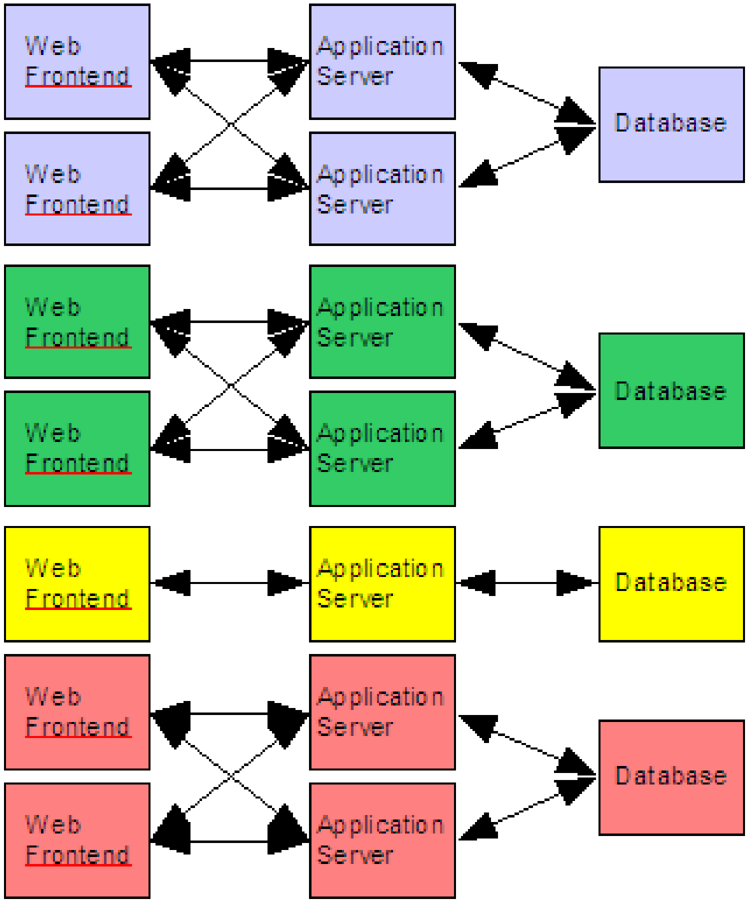
Typical Enterprise Application Architecture

Production plus System Test/QA plus development



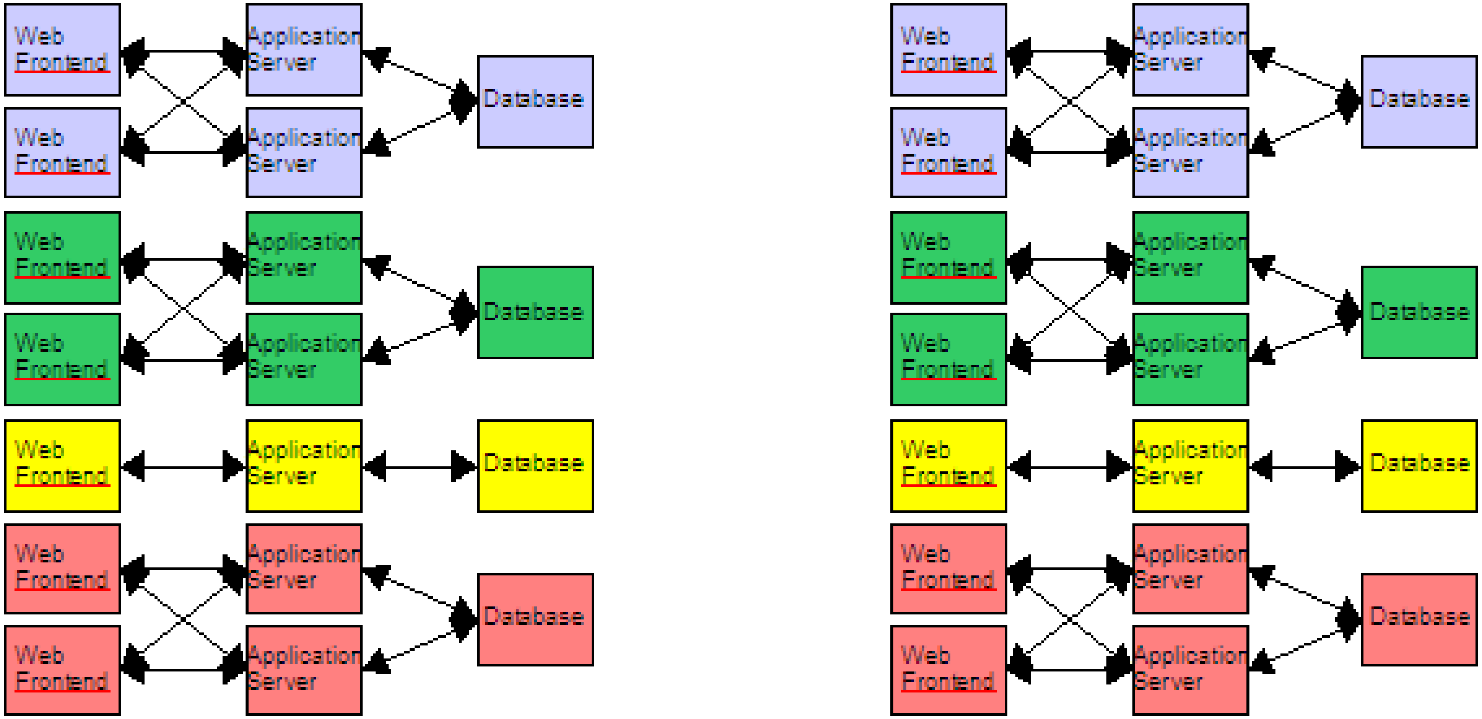
Typical Enterprise Application Architecture

Production plus System Test/QA plus development plus DR



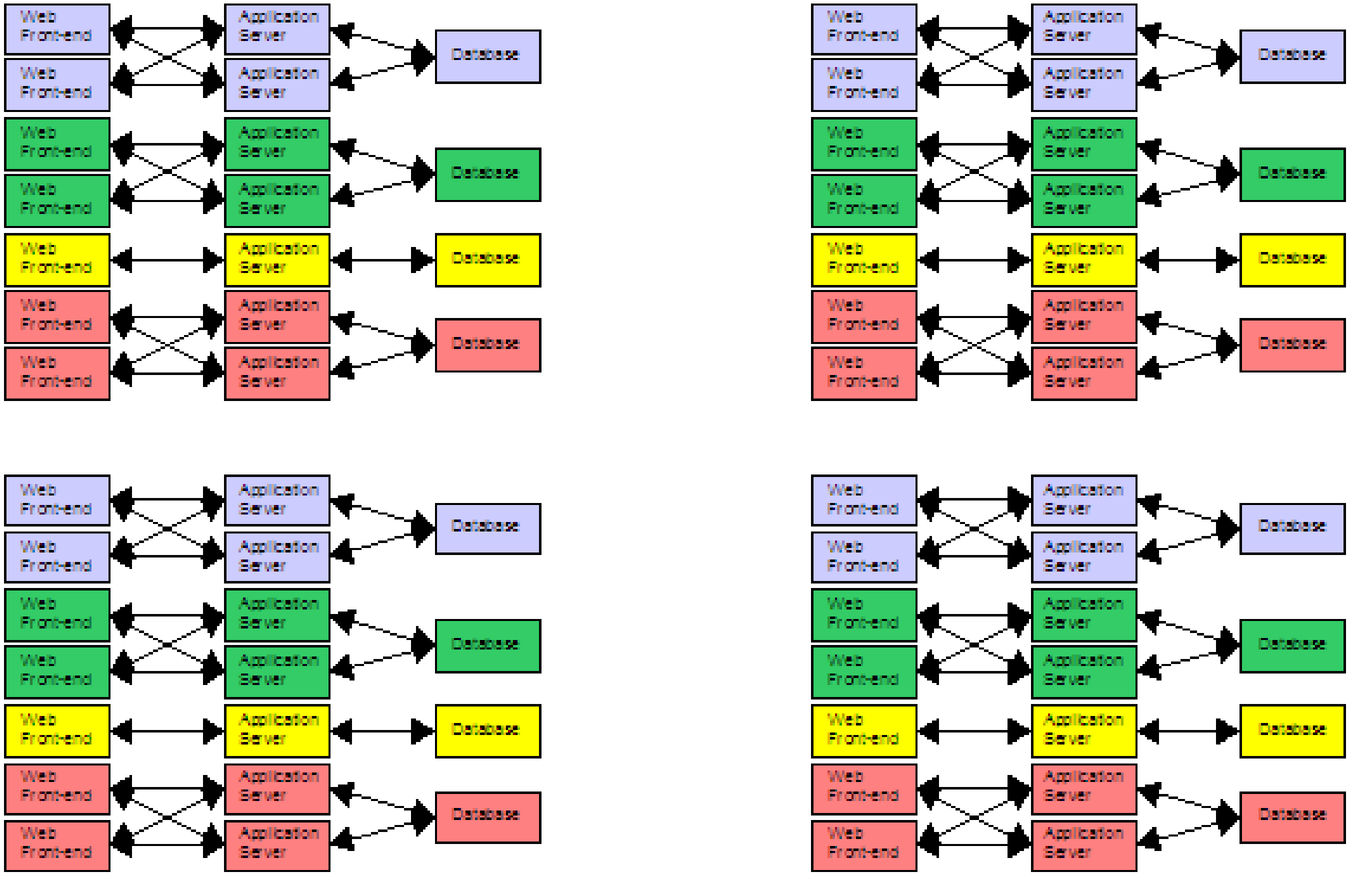
Typical Enterprise Application Architecture

Two applications



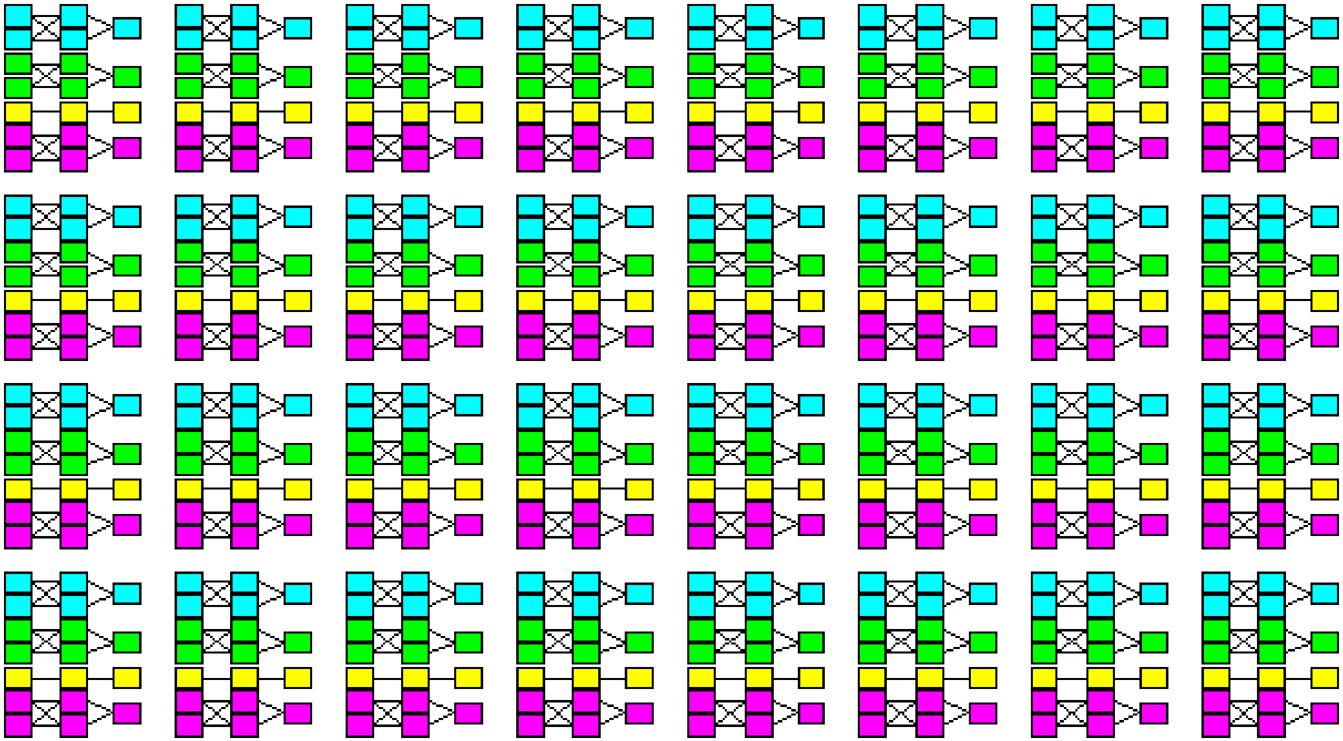
Typical Enterprise Application Architecture

Four applications

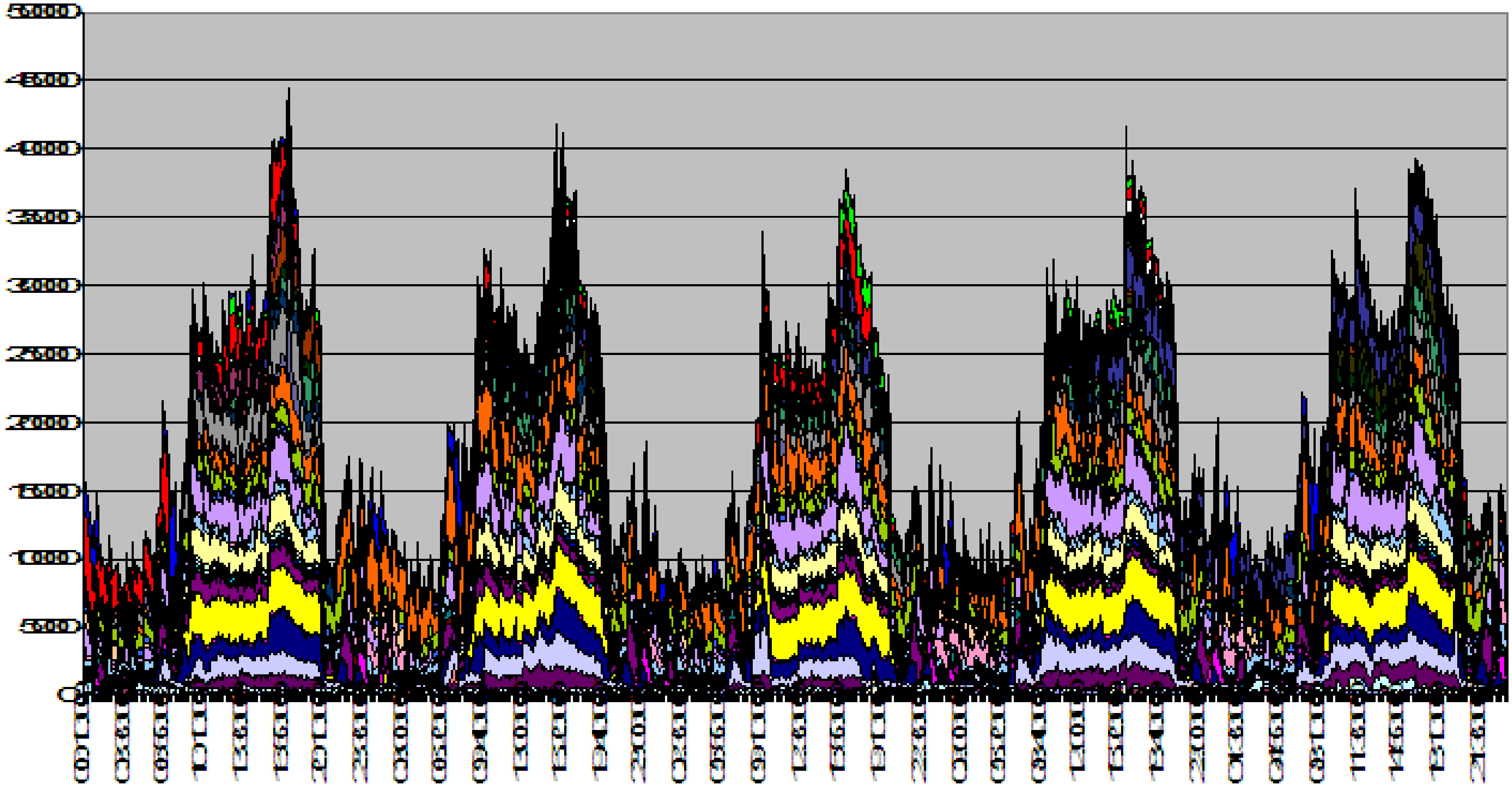


Typical Enterprise Application Architecture

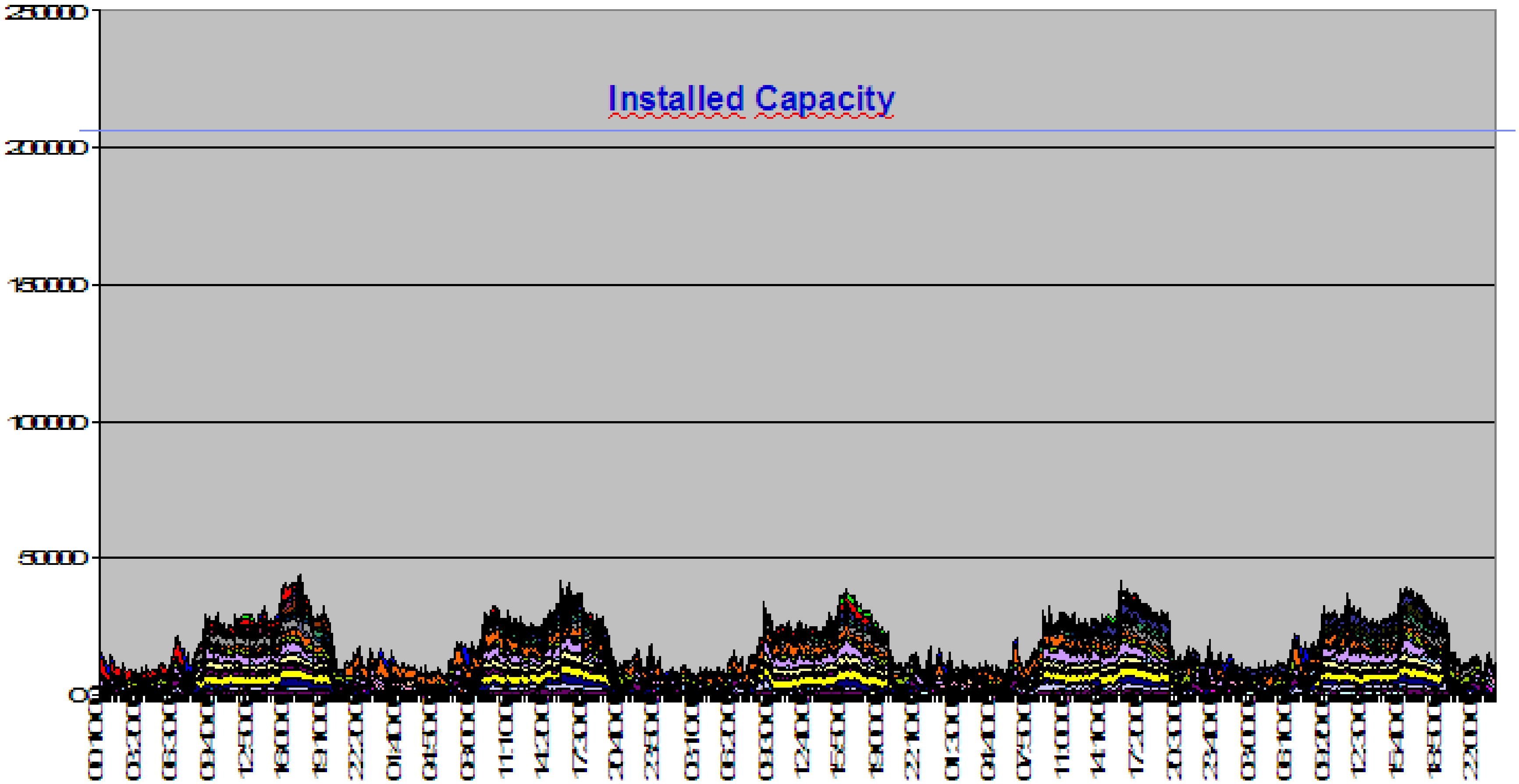
32 applications and we're up to 576 servers
Large enterprises have dozens to hundreds of applications...



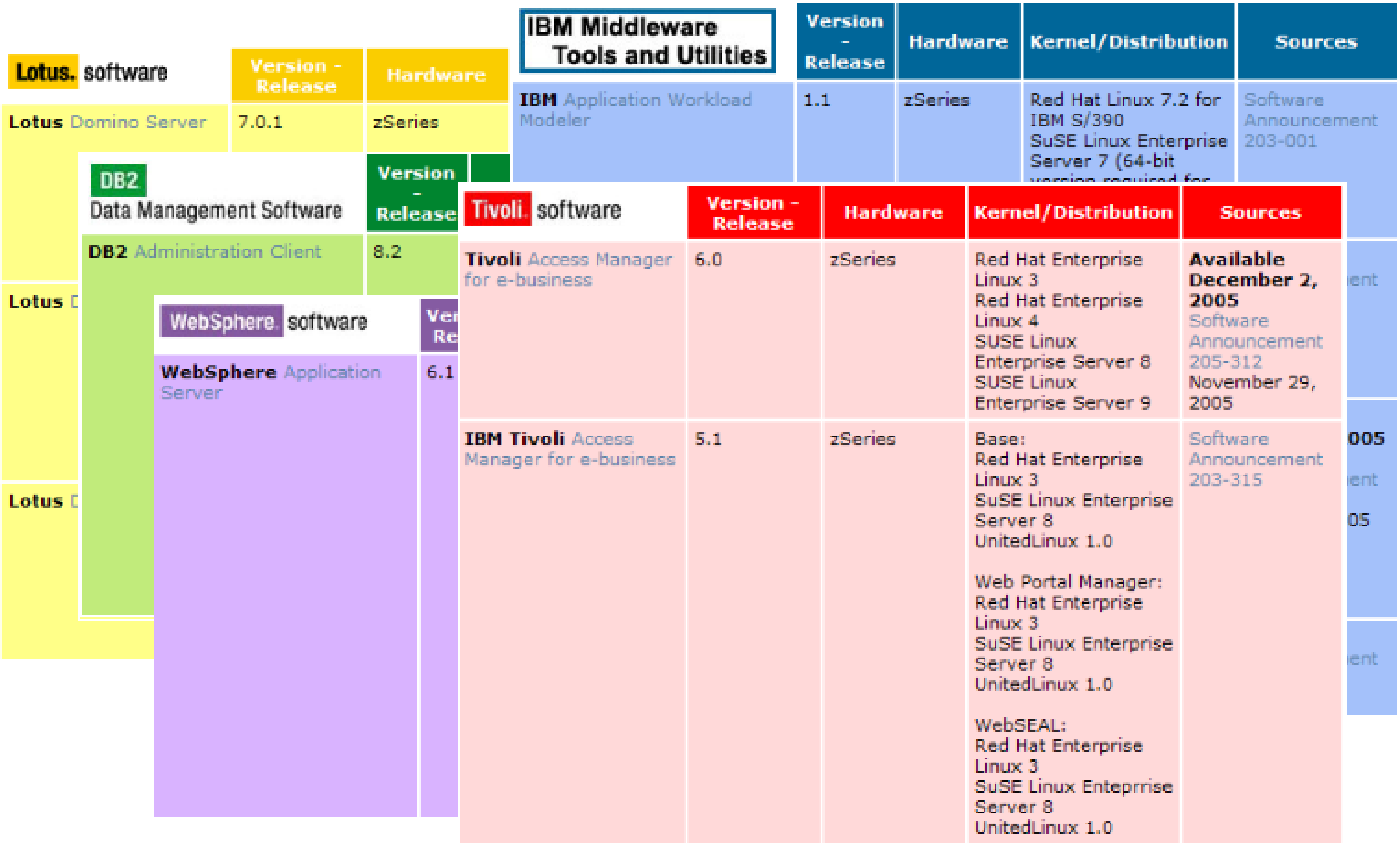
Typical Server Utilisation (absolute)



Typical Server Utilisation (relative)



IBM Software for Linux – Product Matrix *ibm.com/linux/matrix*



Lotus. software	Version - Release	Hardware
Lotus Domino Server	7.0.1	zSeries
Lotus D		
Lotus D		

DB2 Data Management Software	Version - Release
DB2 Administration Client	8.2

WebSphere. software	Ver Re
WebSphere Application Server	6.1

IBM Middleware Tools and Utilities	Version - Release	Hardware	Kernel/Distribution	Sources
IBM Application Workload Modeler	1.1	zSeries	Red Hat Linux 7.2 for IBM S/390 SuSE Linux Enterprise Server 7 (64-bit version required for	Software Announcement 203-001

Tivoli. software	Version - Release	Hardware	Kernel/Distribution	Sources
Tivoli Access Manager for e-business	6.0	zSeries	Red Hat Enterprise Linux 3 Red Hat Enterprise Linux 4 SUSE Linux Enterprise Server 8 SUSE Linux Enterprise Server 9	**Available December 2, 2005** Software Announcement 205-312 November 29, 2005
IBM Tivoli Access Manager for e-business	5.1	zSeries	Base: Red Hat Enterprise Linux 3 SuSE Linux Enterprise Server 8 UnitedLinux 1.0 Web Portal Manager: Red Hat Enterprise Linux 3 SuSE Linux Enterprise Server 8 UnitedLinux 1.0 WebSEAL: Red Hat Enterprise Linux 3 SuSE Linux Enteprrise Server 8 UnitedLinux 1.0	Software Announcement 203-315

ISV Support for Linux on System z

Linux on zSeries

Solutions
Software
Services
Technical support
Education

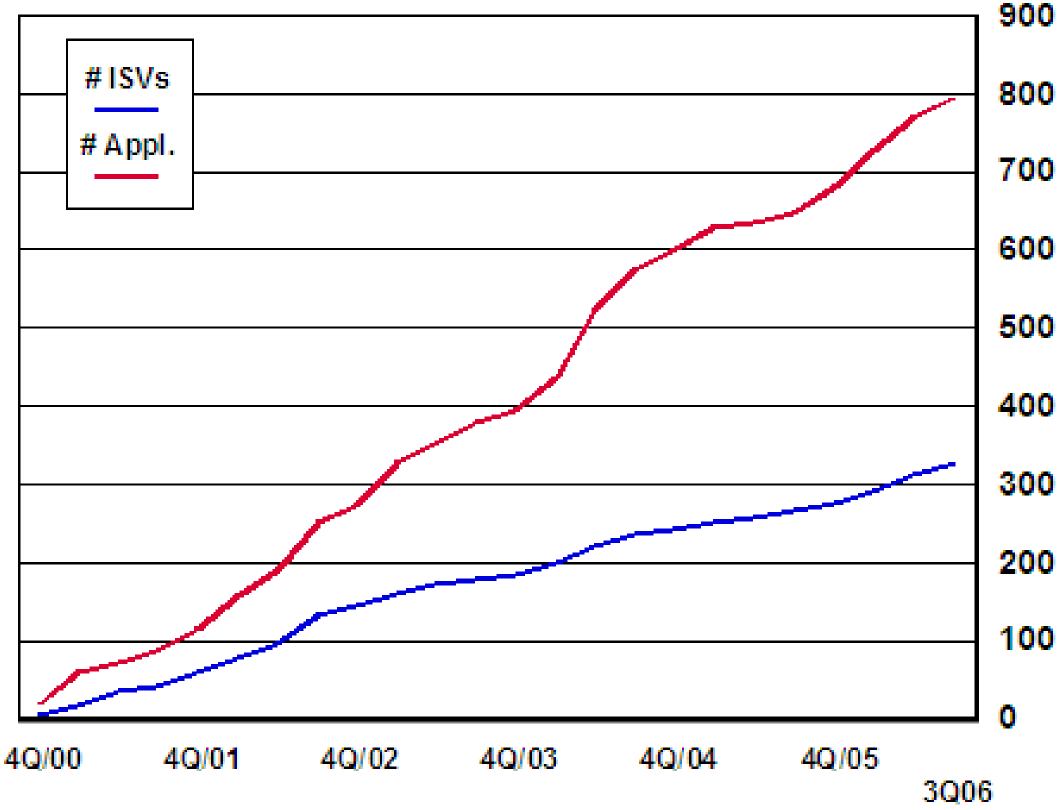
Available applications

All

Company	Application	Category	Availability

Views

By category

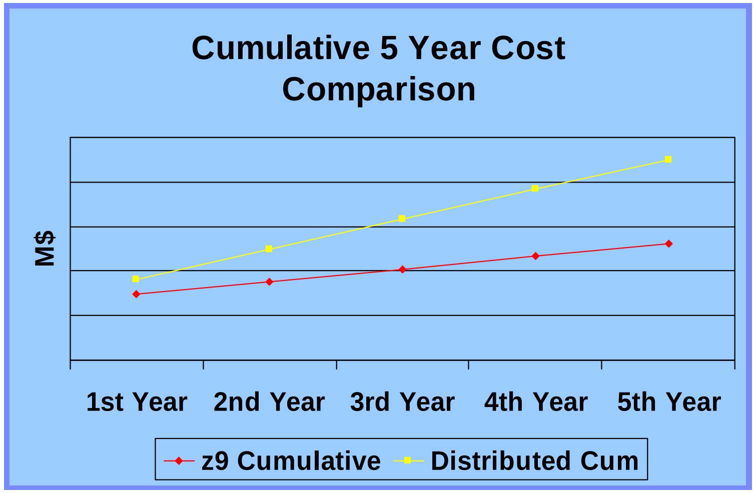


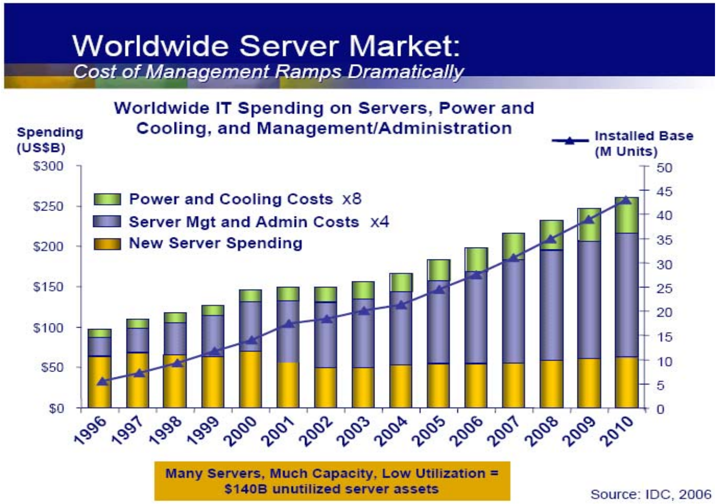
For more details please see:
http://www.ibm.com/servers/eserver/zseries/os/linux/apps/all.html

IBM Internal Linux Consolidation

- *IBM will consolidate about 3,900 servers onto about 30 System z mainframes running Linux*
- Used commercial TCO model to estimate savings in a Cross-IBM effort.
- *We expect substantial savings :*
  - Annual Energy Usage reduced by 80%
  - Total floor space reduced by 85%
- *This transformation is enabled by the System z's sophisticated*

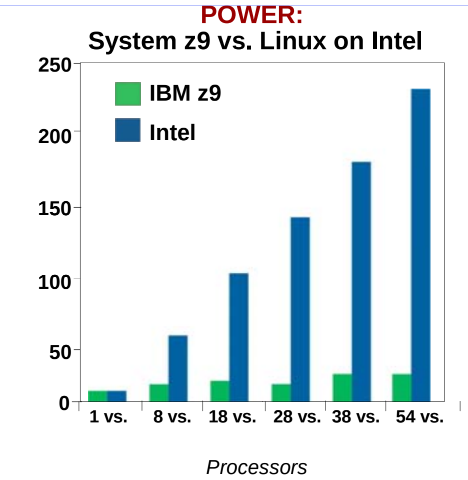
  virtualisation capability

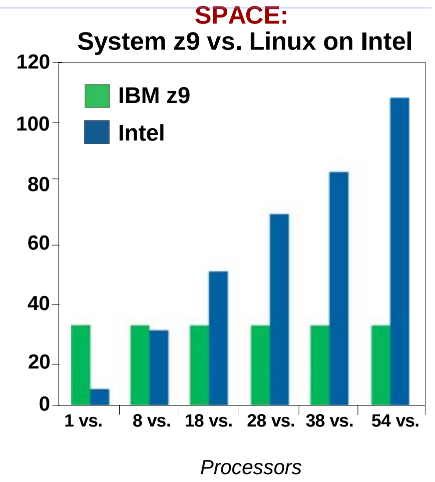




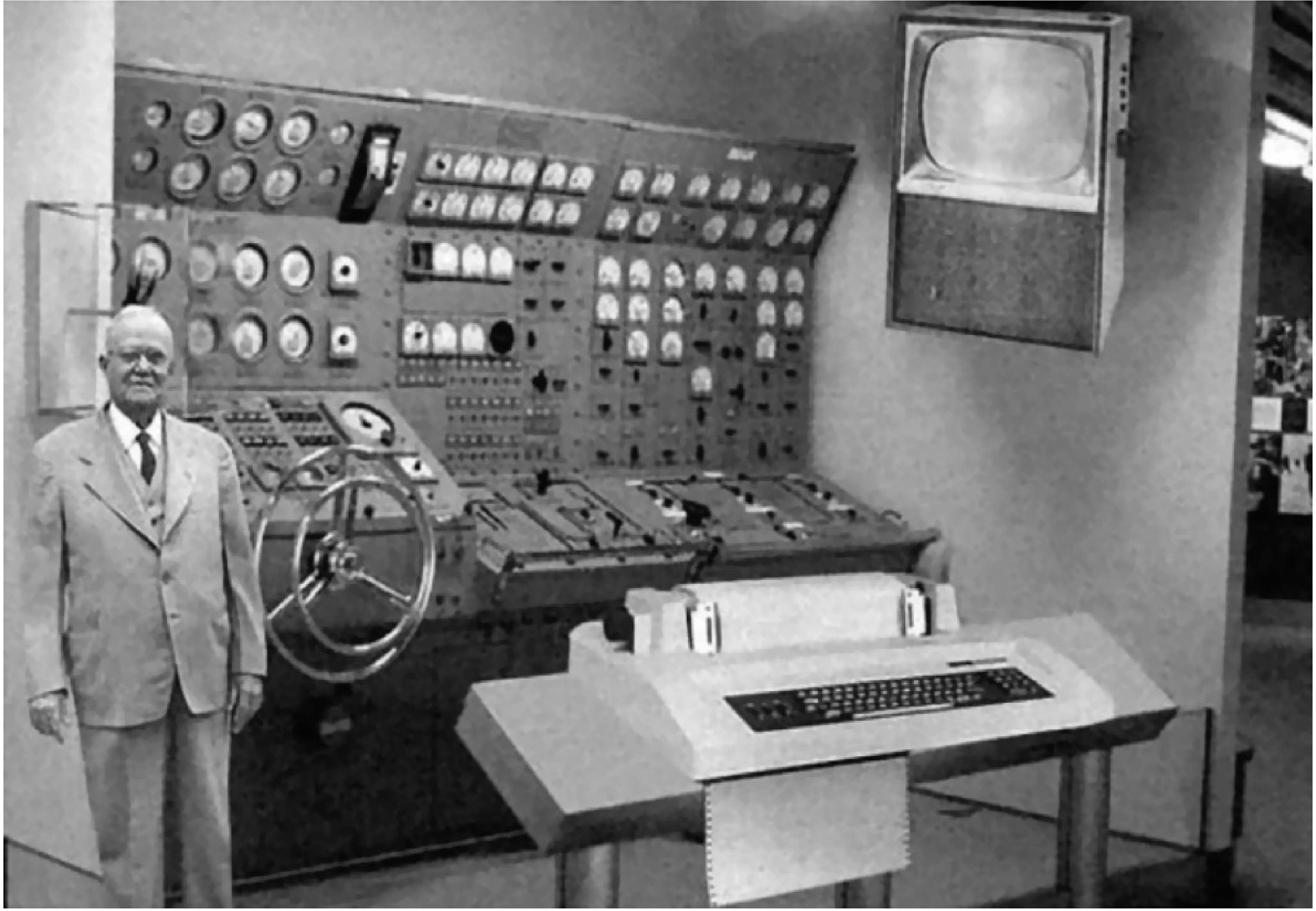
Many Servers, Much Capacity, Low Utilization = $140B unutilized server assets

Source: IDC, 2006





Source: IBM. The Linux on Intel servers selected in this example are functionally eligible servers considered for



Scientists from the RAND Corporation have created this model to illustrate how a 'home computer' Could look like in the year 2004. However the needed technology will not be economically feasible for the average home. Also the scientists readily admit that the computer will require not yet invented Technology to actually work, but 50 years from now scientific progress is expected to solve these problems with teletype interface and the Fortram language, the computer will be easy to use.

Thankyou for your attention