z/VM et Linux for zSeries

Jerome LE – IBM GTS
Spécialiste z/VM et Linux for zSeries

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX*	IBM*	PowerVM	System z10	z/OS*
BladeCenter*	IBM eServer	PR/SM	WebSphere*	zSeries*
DataPower*	IBM (logo)*	Smarter Planet	z9*	z/VM*
DB2*	InfiniBand*	System x*	z10 BC	z/VSE
FICON*	Parallel Sysplex*	System z*	z10 EC	
GDPS*	POWER*	System z9*	zEnterprise	
HiperSockets	POWER7*			

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Windows Server and the Windows logo are trademarks of the Microsoft group of countries.
InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Agenda

- Chapitre 1 - Notions de base
  - Data Center
  - Consolidation de serveurs
  - Les techniques de Virtualisation
  - Virtualisation pour xSeries
  - Virtualisation pour Power Architecture
  - Virtualisation avec Linux
  - Le zEnterprise System
  - Les disques
  - Les cartes OSA
  - Les bandes
- Chapitre 2 – Virtualisation pour le zEnterprise
  - LPAR
  - z/VM
  - Notions de base z/VM
  - Le Réseau
  - Installation de z/VM
  - Performance Toolkit
  - DIRMAINT
  - RACF
  - RSCS
  - Autres produits
- Chapitre 3 – Linux for zSeries
  - L'Open
  - Consolidation
  - La mise en oeuvre
  - L'accès à distance
  - Administration de Linux
  - Le clonage
  - Les sauvegardes
  - Le réseau
  - De l'information

Chapitre 1- Notions de base

Data center

Les rôles

Application Developer

Roles

Production Control Analyst

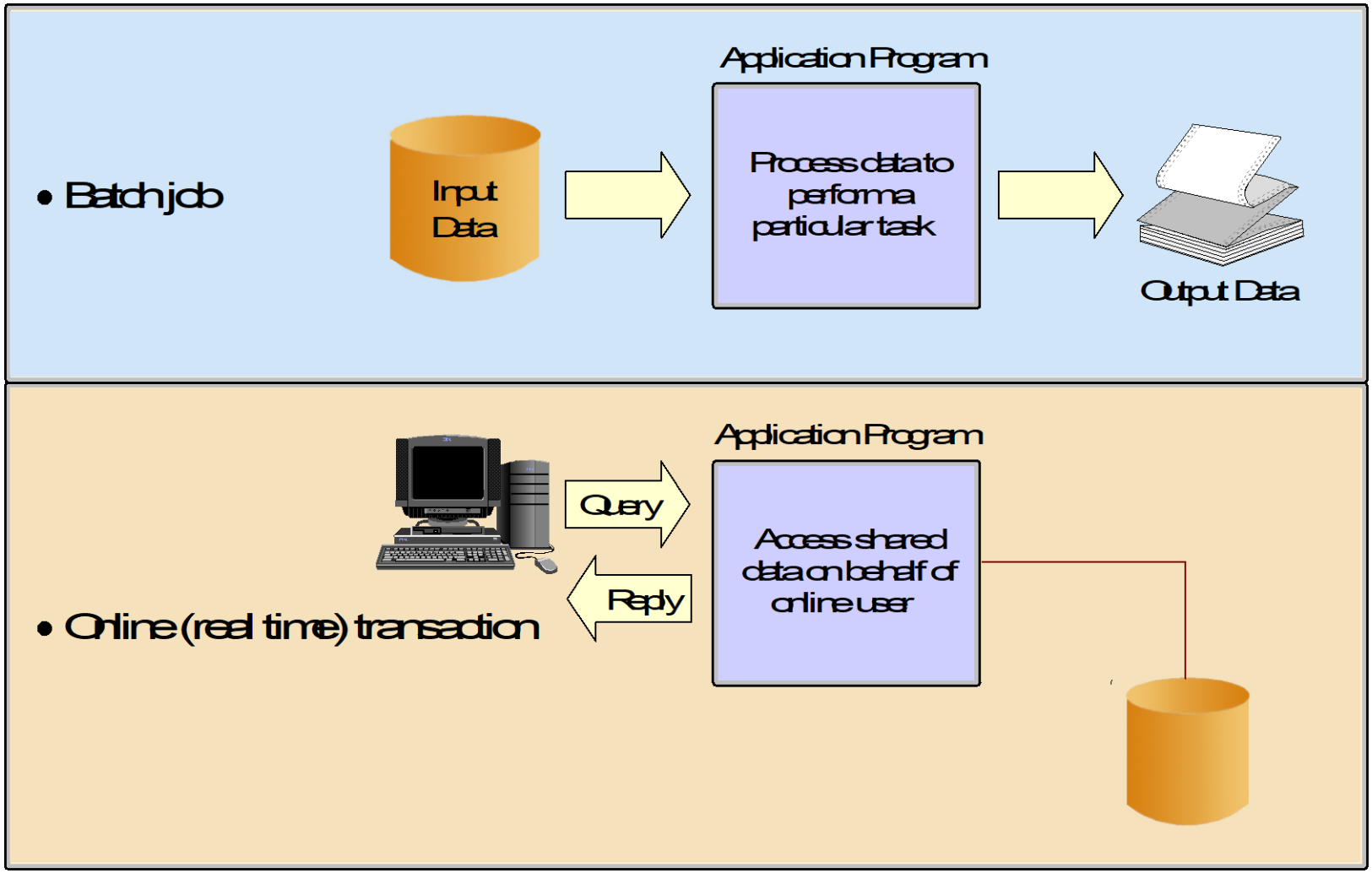
End User

Operator

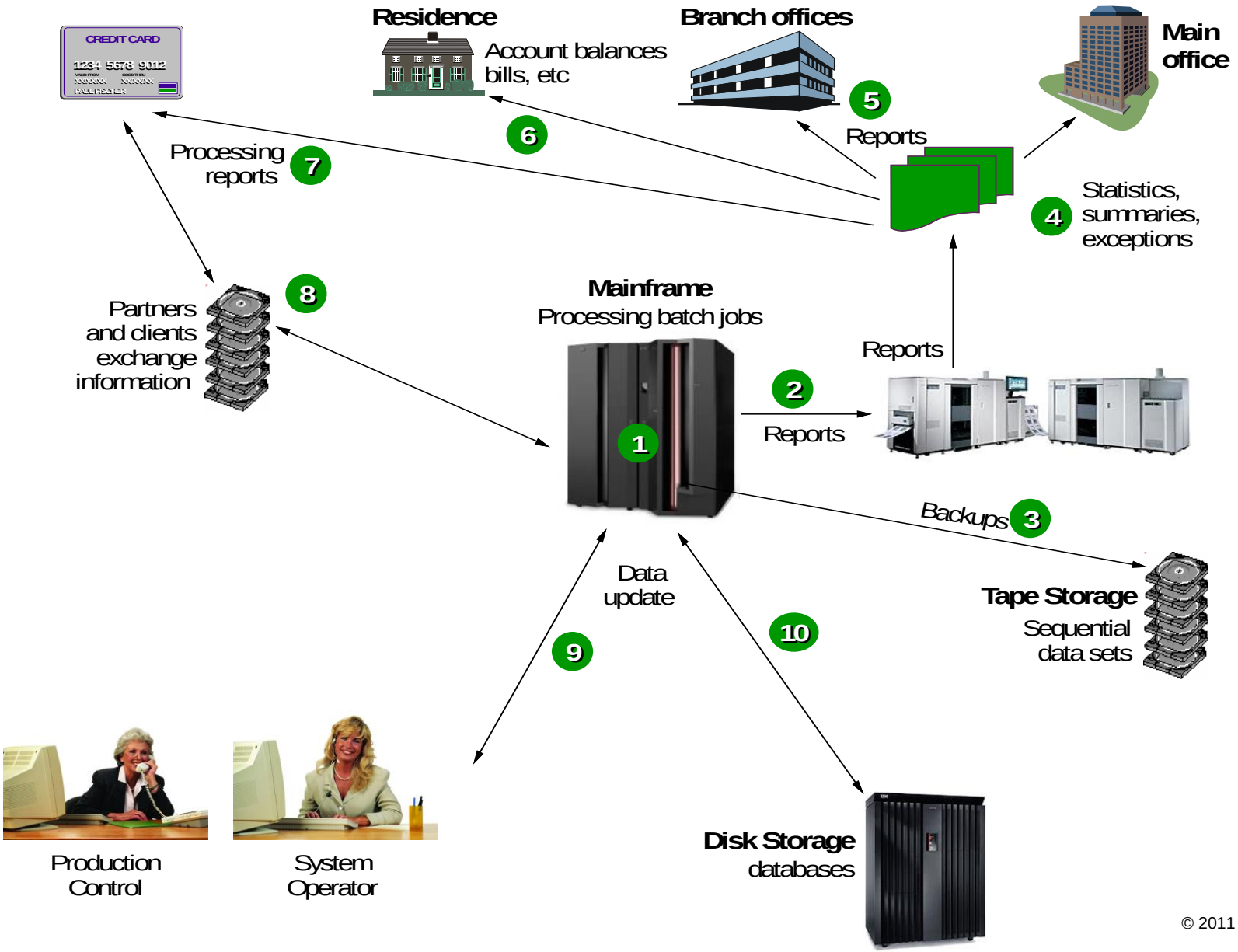
System Programmer

System Administrator

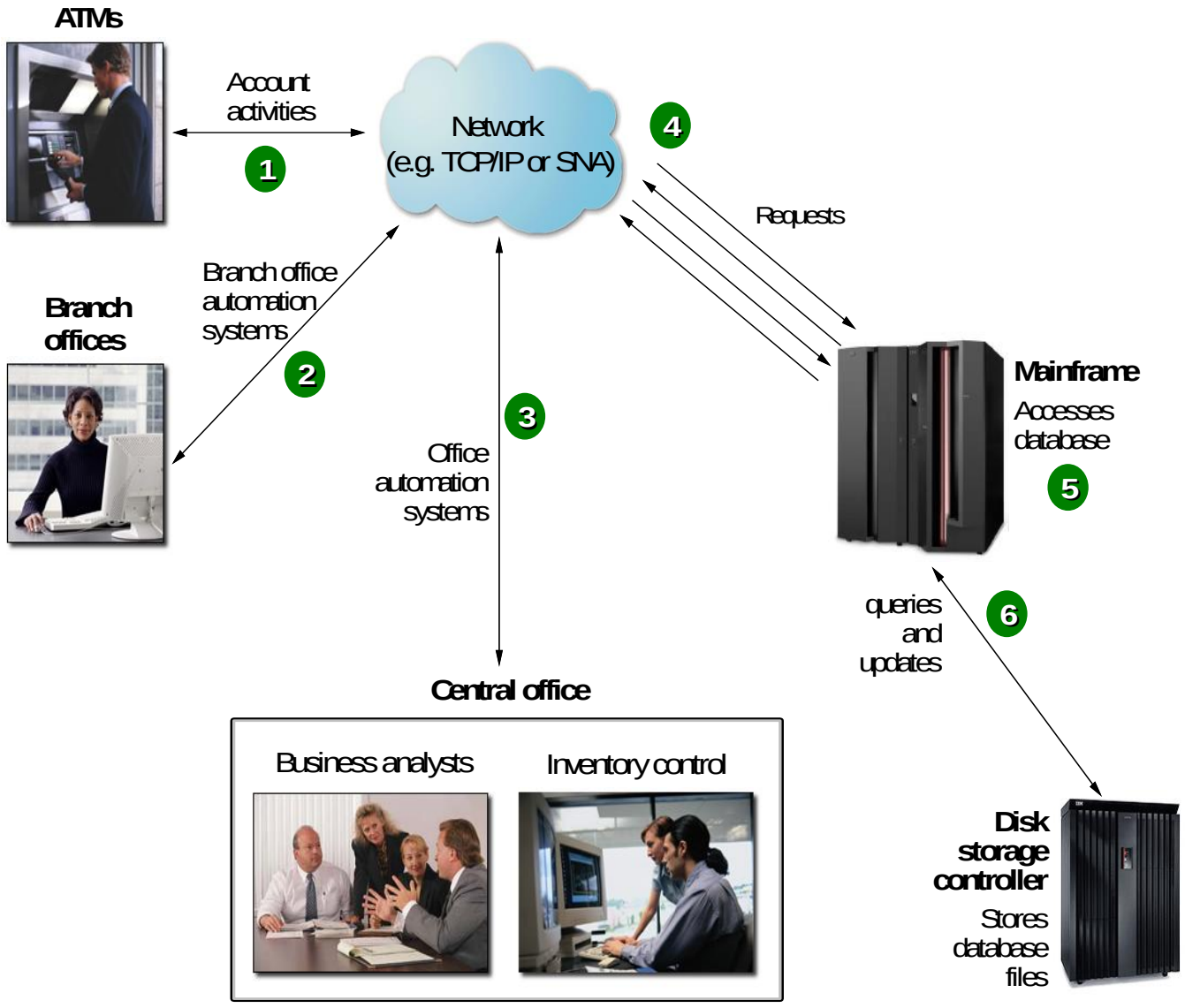
Types de traitements d'un Centre Informatique



Traitement batch



Les transactions



Le support matériel et logiciel

- La licence (droit d'utilisation du logiciel) en zSeries est basée sur le nombre de processeurs ou sur la consommation processeur du produit.
- Le support des produits IBM

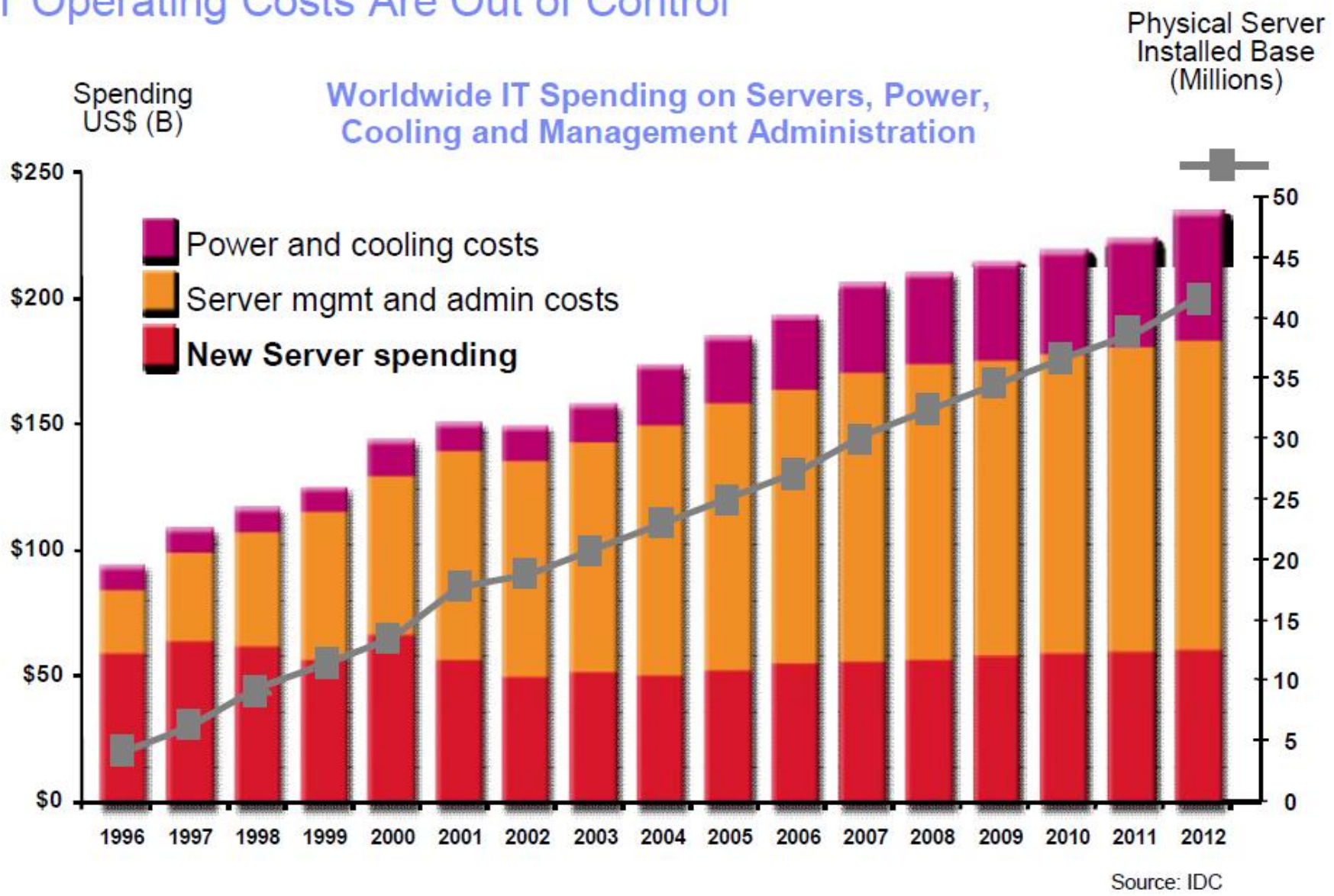
  – Matériel
    - Contrat de maintenance
    - Le contrat de maintenance couvre aussi l'aide à la détermination de l'origine du problème (matériel ou logiciel)

  – Logiciel
    - La licence donne accès aux corrections existantes et à la correction de défaut. L'analyse du problème étant décrite dans un APAR (Authorized Program Analysis Report)
    - L'offre Point Service aide à l'analyse (dump, trace) des problèmes. Elle assure aussi la réponse aux questions de paramétrage ou de mise en oeuvre.
    - L'offre ETS (Extended Technical Support) apporte une relation personnalisée pour la gestion et le suivi des problèmes rencontrés. Une approche proactive pour anticiper et éviter des incidents.

Consolidation de serveurs

Le cout de la multiplication des serveurs

IT Operating Costs Are Out of Control



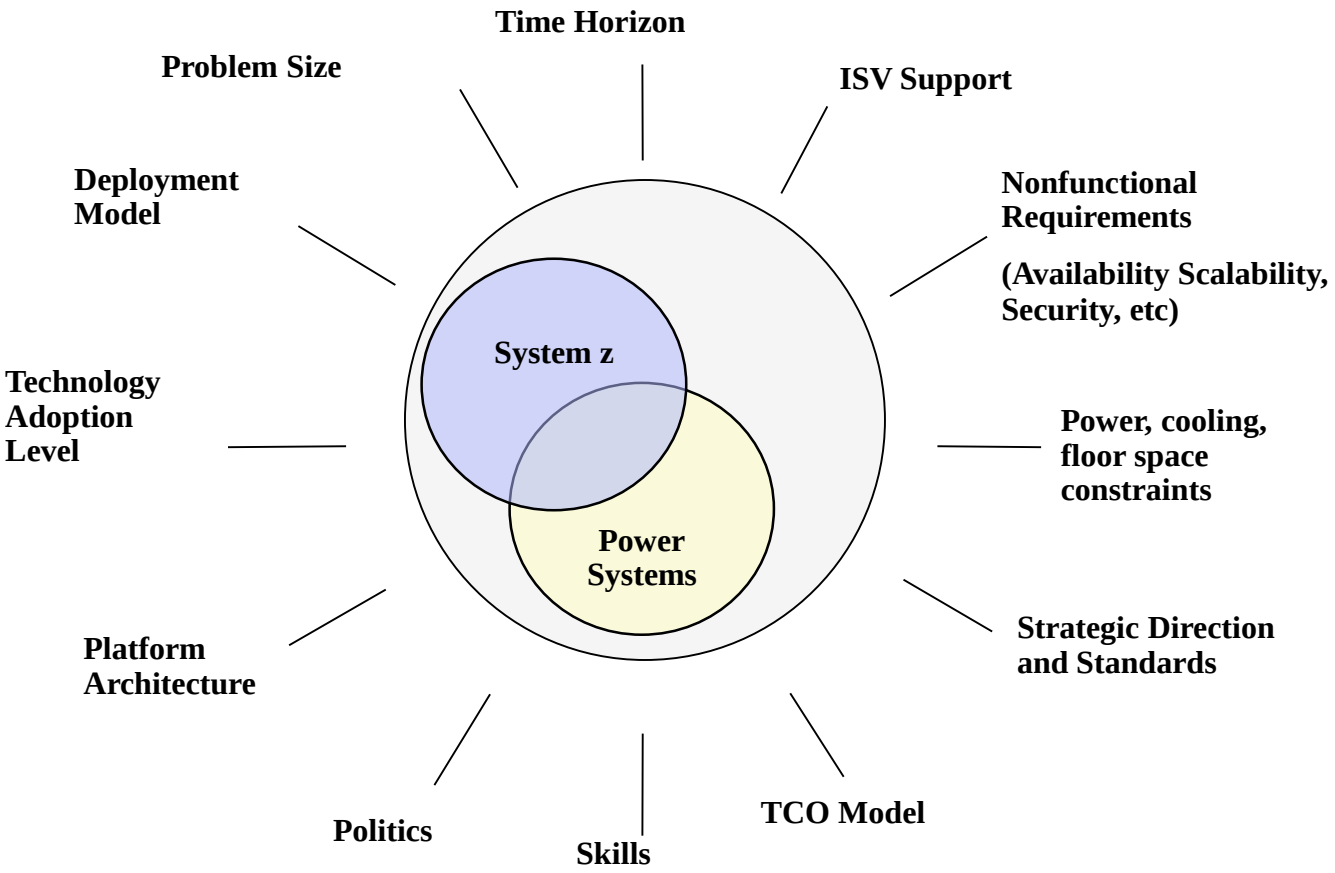
Consolidation

- Regroupement de services sur une même infrastructure afin
  - D'optimiser l'utilisation des ressources et des couts
  - D'améliorer la fiabilité de l'ensemble
- L'étude de TCO (Total Cost of Ownership) décrit les couts globaux d'une solution sur 3 années
  - Matériels (Serveurs, Disques, Réseaux, System Management, Racks + cable)
  - Logiciels (Operating System, System management, Database, Applications, Support)
  - Personnels en FTE (Full Time Equivalent)
  - Environnement - Occupation (place au sol), Energies (Electricité, Refroidissement)
  - Migration
  - Couts de l'indisponibilité
- L'étude de TCA (Total Cost of Acquisition) recense les frais pour la mise en œuvre de la solution
  - Matériels
  - Logiciels
  - Formations

Les paramètres d'une étude d'architecture

- La scalabilité (capacité à faire évoluer la puissance de traitement, à absorber des pics de production, à faire évoluer les ressources rapidement)
- L'infrastructure réseau à mettre en œuvre pour la solution
- La qualité de support de chaque élément (matériels, logiciels)
- L'optimisation du partage des ressources et les limites
- Les logiciels nécessaires et leurs limites
- Les besoins en énergie et en surface
- Les niveaux de sécurité à assurer
- L'expérience acquise de l'entreprise sur ces logiciels ou matériels
- L'importance du PRA (Plan de Reprise d'Activité)
- La qualité de service attendue de la solution

Le diagramme de choix



Le poids de chacun des facteurs décisionnels varie à chaque projet de consolidation.

Real customer examples with real workloads!

Industry	Distributed Cores	IBM Enterprise Linux Server™ Cores	Core-to-Core Ratio*
Public	292	5	58 to 1
Banking	111	4	27 to 1
Finance	442	16	27 to 1
Banking	131	5	26 to 1
Insurance	350	15	23 to 1
Insurance	500+	22	22 to 1
Banking	63	3	21 to 1
Finance	854	53	16 to 1
Health care	144	14	10 to 1
Transportation	84	9	9 to 1
Insurance	7	1	7 to 1

* Client results will vary based on each specific customer environment including types of workloads, utilization levels, target consolidation hardware, and other implementation requirements.

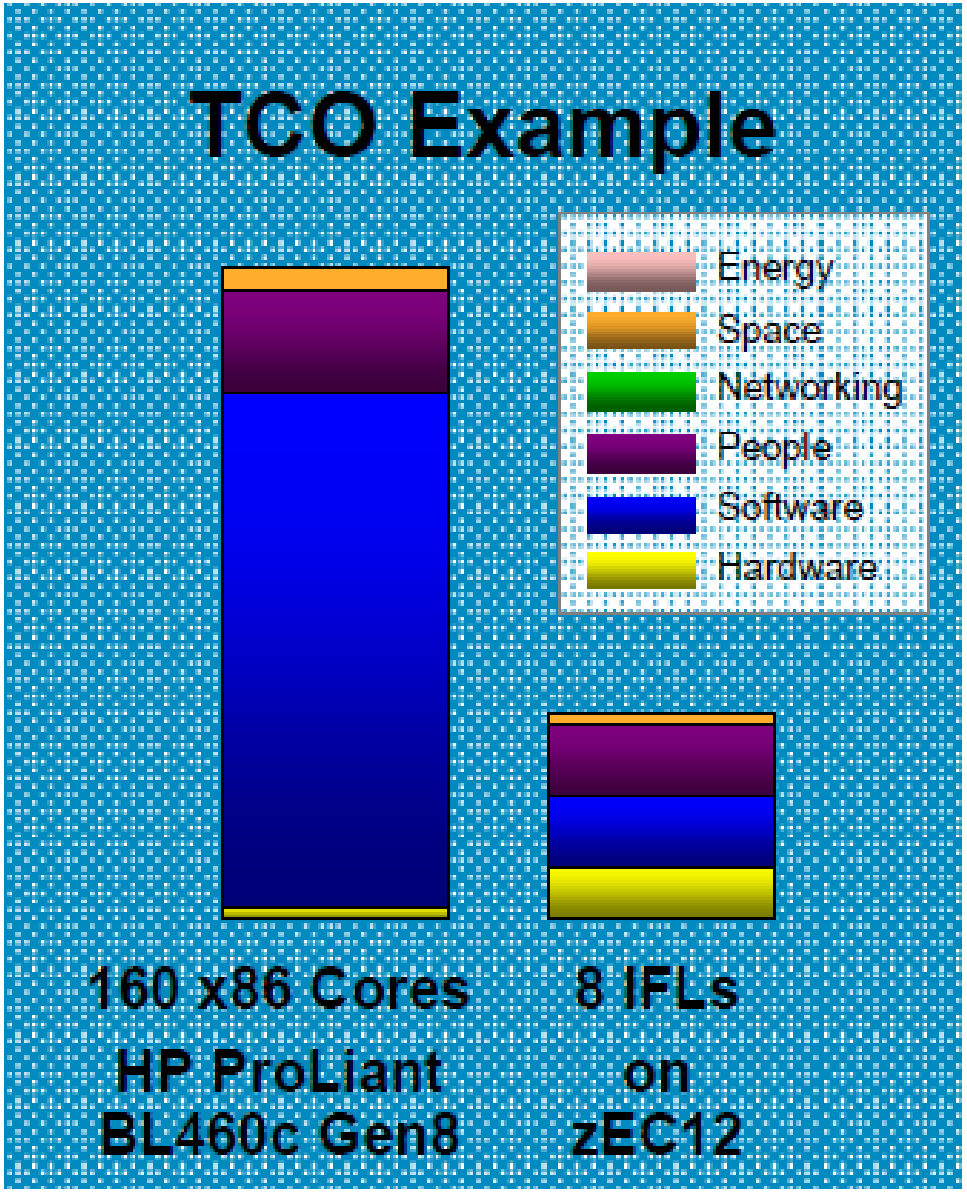
Les conclusions d'un TCO

Linux on System z enables a total cost of acquisition of less than

70 cents per day per virtual server[1]

Consolidate up to 60 distributed cores or more on a single System z core, or thousands on a single footprint[1].

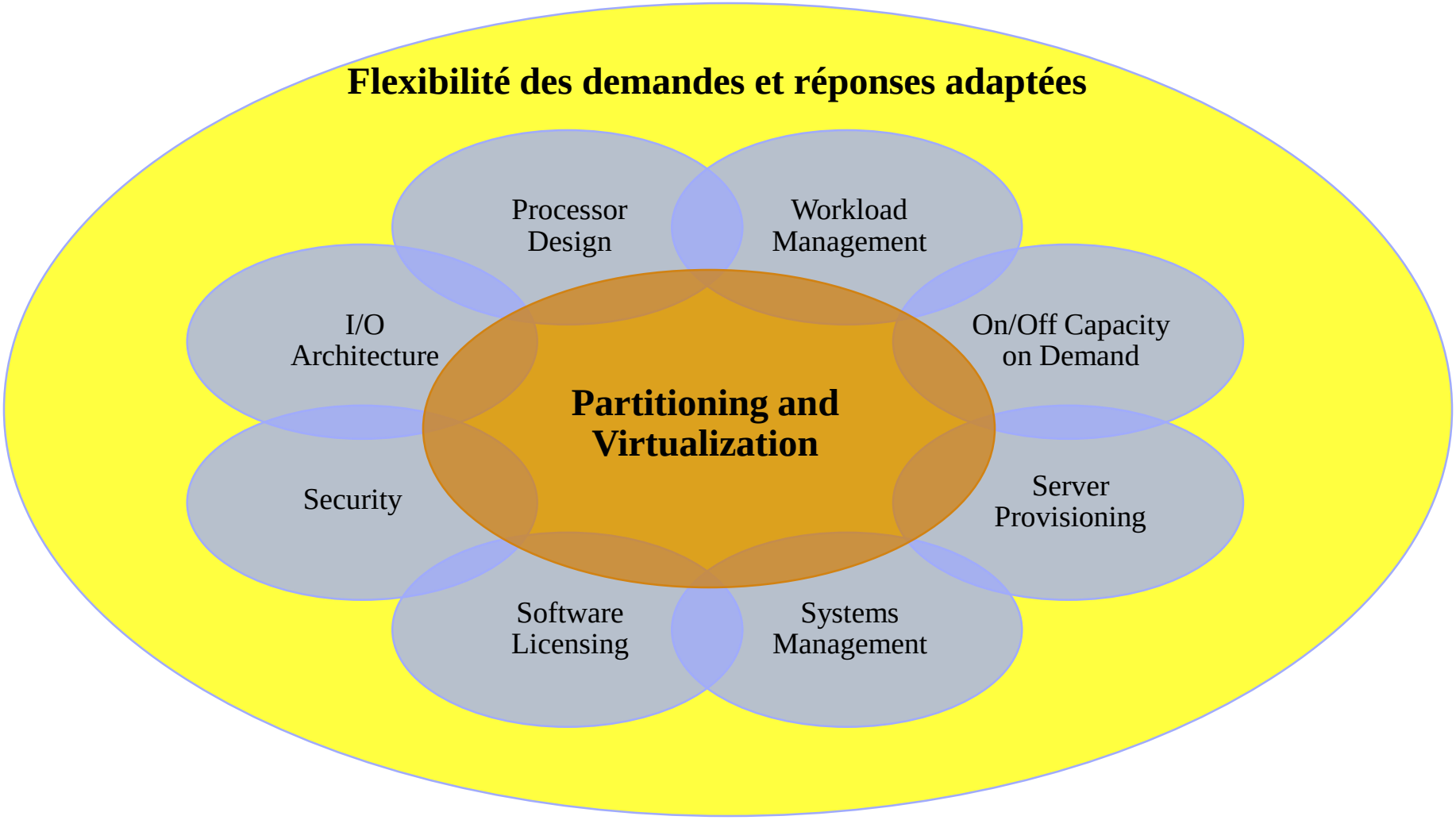
System z servers often run consistently at **90%+ utilization**[1]



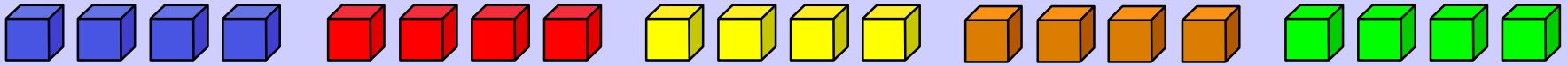
[1] IBM calculations of zEnterprise limits across maximum z196 configuration. Results may vary. 5-Year Total IT Cost

Les techniques de virtualisation

Nécessaire pour l'optimisation des ressources et le déploiement



Concept de la virtualisation



Ressources Virtuelles

- ☒Proxies pour les ressources réelles: mêmes interfaces/fonctions, attributs.
- ☒Parties d'une ressource physique ou des multiples ressources physiques

Virtualisation

- ☒Création des ressources virtuelles et support «map »sur les ressources réelles
- ☒Réalisé avec des Logiciels ou des Fonctions Matérielles

Ressources réelles

- ☒Composants de l'Architecture interfaces/fonctions.
- ☒Peut être centralisé ou distribué.
- ☒Exemples: mémoire, disques, réseaux, serveurs.

L’histoire de la virtualisation

1967 IBM

- CP67
- Architecture 360
- Time Sharing

1972 IBM

- Virtual Machine
- DAT
- Mémoire Virtuelle

1974 Communication de Goldberg et Popek ont défini :

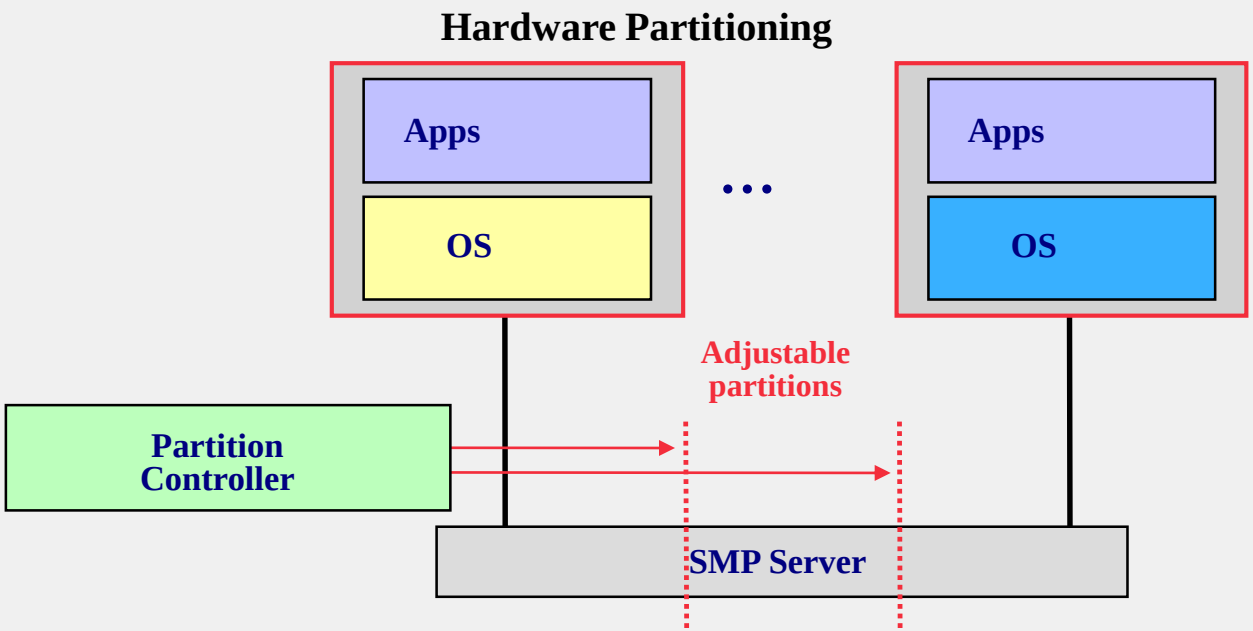
- VMM (Virtual Machine Monitor) ⇔ Hyperviseur
- L’Equivalence : fonctionne à l'identique en virtuel ou en réel
- Le contrôle des ressources : le VMM assure le contrôle
- L’efficacité : la majeur partie des instructions doivent ne pas être dépendante du VMM
- Les instructions du processeur doivent être de 3 types

  Privilégiées (Mode hyperviseur)

  Sensitives (Modification de configuration) sous ensemble des privilégiés

  Non-Privilégiées (exécutées directement par l’OS virtualisé)

Le partitionnement matériel



Server is subdivided into fractions each of which can run an OS

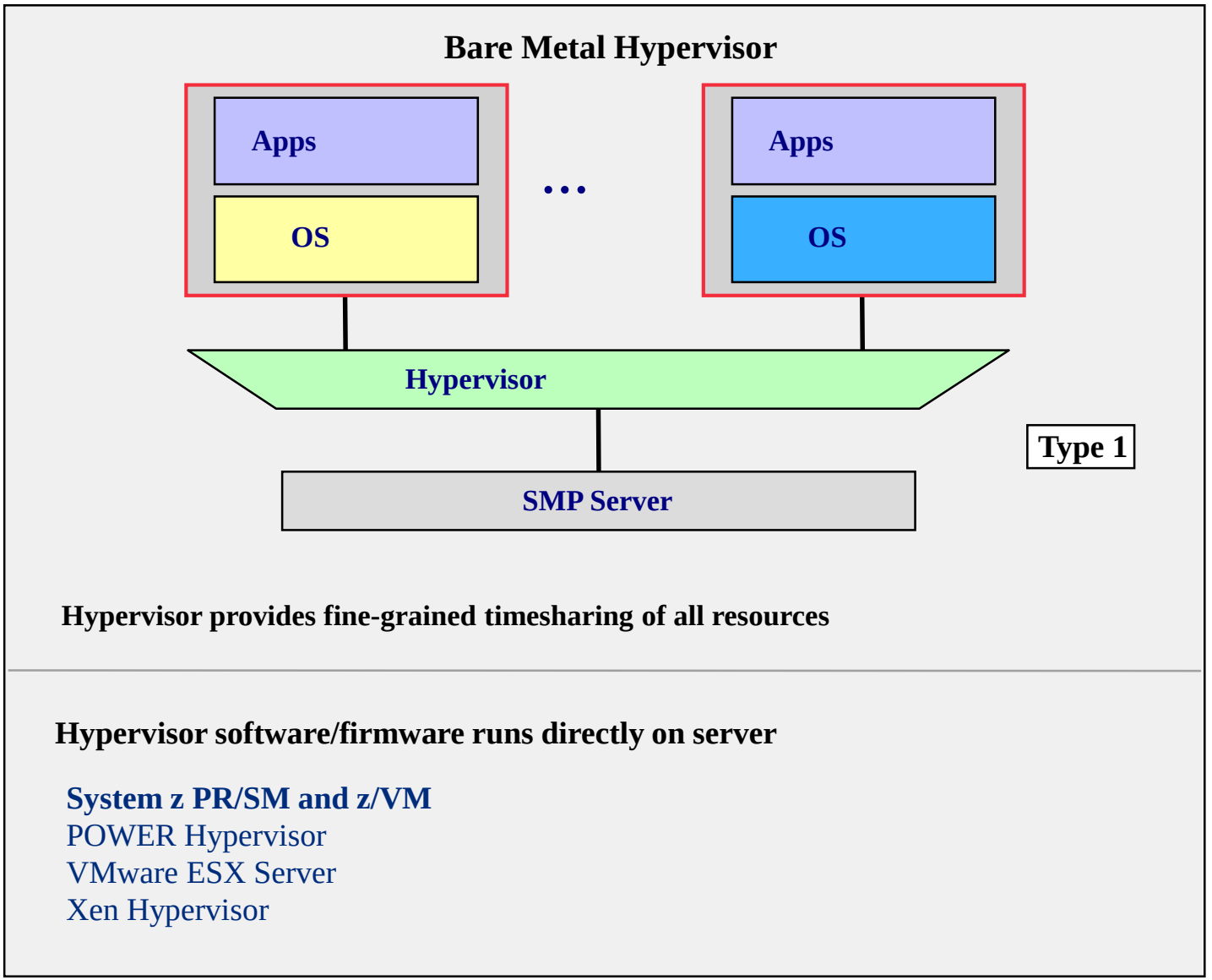
Physical partitioning

Sun Domains, HP nPartitions

Logical partitioning

pSeries LPAR, HP (PA) vPartitions

Hyperviseur de type 1



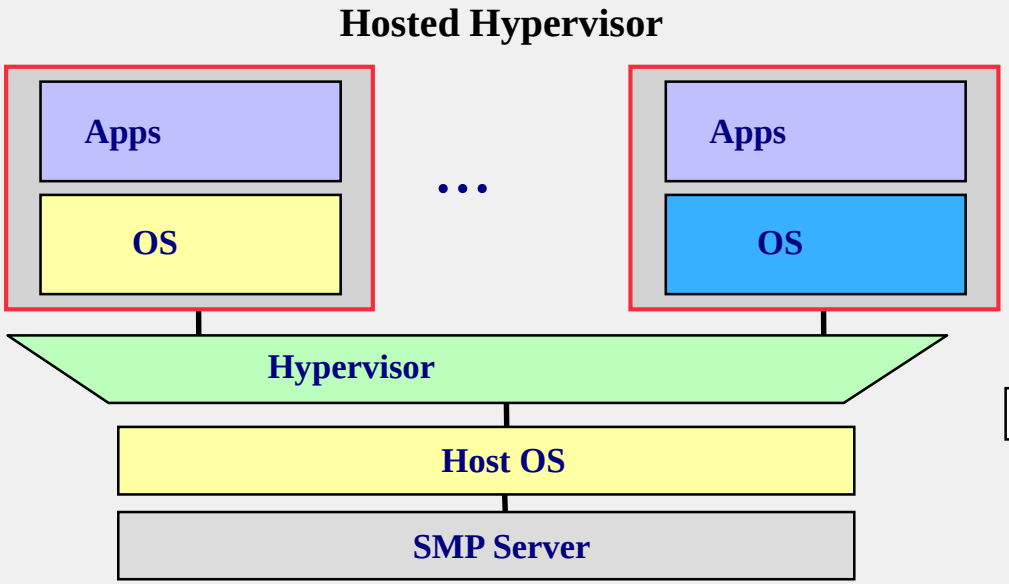
Hypervisor provides fine-grained timesharing of all resources

Hypervisor software/firmware runs directly on server

System z PR/SM and z/VM
POWER Hypervisor
VMware ESX Server
Xen Hypervisor

PR/SM = Processor Resource/System Manager

Hyperviseur de type 2

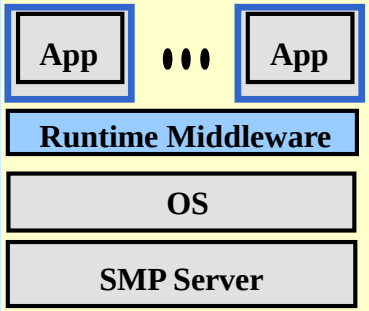
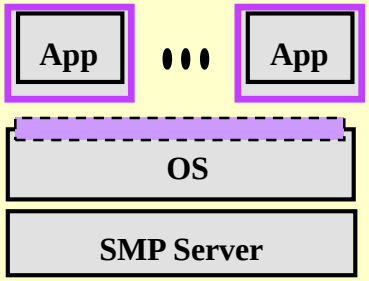


Hypervisor uses OS services to do timesharing of all ressources

Hypervisor software runs on a host operating system

VMware GSX
Microsoft Virtual Server
HP Integrity VM
User Mode Linux

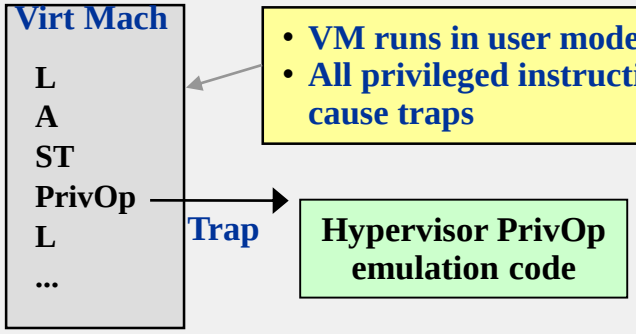
Virtualisation des workloads

	Method	Description	Examples
Application Containers	 **Virtual Runtimes**	▪ Middleware provides JVM, J2EE, or CLR container per application ▪ Virtual runtimes can be OS independent	▪ WebSphere XD ▪ ….
	 **Virtual OS Environ-ments.**	▪ OS and middleware creates virtual OS environment per application ▪ Each container has its own name space, files, root, …	▪ z/OS Ad. Sp., i5/OS Subsys. ▪ Solaris Containers ▪ AIX (to be announced) ▪ HP-UX Secure Resource Part. ▪ Windows & Linux offerings from Softricity, SWsoft, VERITAS, Trigence, …

This is more Workload Management than it is virtualization

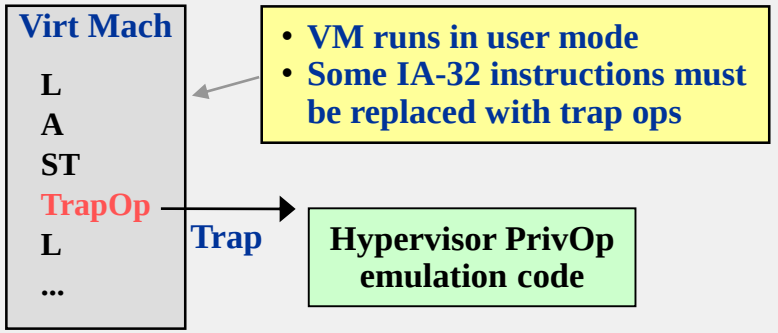
La détection des instructions privilégiées par l'Hyperviseur

Trap and Emulate



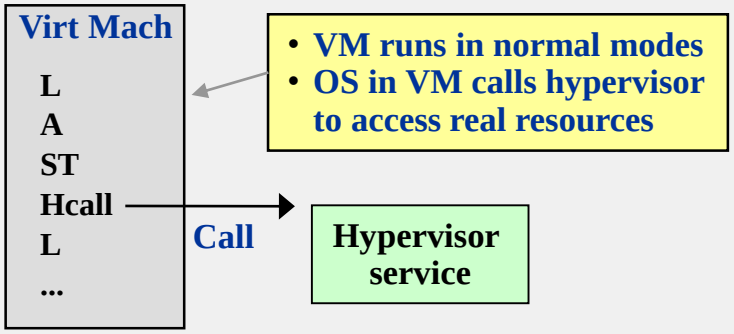
Examples: CP-67, VM/370
Benefits: Runs unmodified OS
Issues: Substantial overhead

Translate, Trap, and Emulate



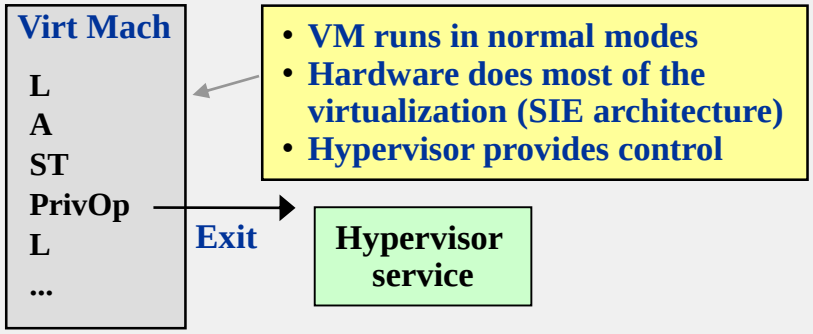
Examples: VMware, Microsoft VS
Benefits: Runs unmodified, translated OS
Issues: Substantial overhead

Hypervisor Calls ("Paravirtualization")



Examples: POWER Hypervisor, Xen, CMS
Benefits: High efficiency
Issues: OS must be modified to issue Hcalls

Direct Hardware Virtualization



Examples: **PR/SM, z/VM** (also use hypervisor calls)
Benefits: Highest efficiency. Runs unmodified OS
Issues: Requires underlying HW support

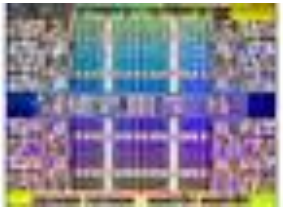
Les hyperviseurs pour Intel/AMD

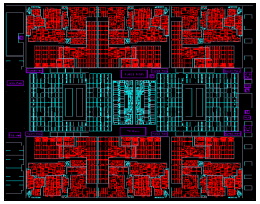
- Utilisé pour virtualiser de 2 à 30 serveurs
- VMware (EMC)
  - Hyperviseur de type 1
  - Maximum 32 vCPUs et 96Go de mémoire
  - Windows, Linux, Solaris
- Xen (CITRIX)
  - Hyperviseur de type 1
  - Maximum 8vCPUs et 256Go de mémoire
  - Windows, Linux, Solaris
- Hyper-V (Microsoft)
  - Hyperviseur de type 2
  - Maximum 64pCPU, 4vCPU
  - Mémoire 1To
  - Windows, Linux

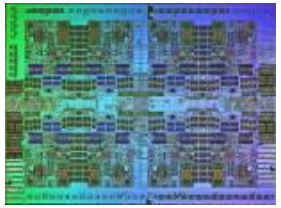
Tableau comparatif

Scalability Factors	VMware ESXi 5 *(EMC)*	HyperV (Microsoft)	XEN *(Citrix)*	PowerVM (IBM)	z/VM (IBM)
Processor	Intel	Intel	Intel	Power	zProcessor
Hyperviseur Type	1	2	1	1	1
Virtual CPUs per VM	32	4	16	256	64
Memory per VM	96 GB	64 GB	128 GB	8192 GB	16 EB
Live VMs per server	512	384	512	1000	8 chars name > 1000000
Memory per server	2 TB	1TB	1 TB	8 TB	10 TB
OperatingSystem	Windows, Linux	Windows, Linux	Windows, Linux, Solaris	AIX, IBM I, Linux	z/OS, z/VM, z/VSE, z/TPF, Linux

Les processeurs



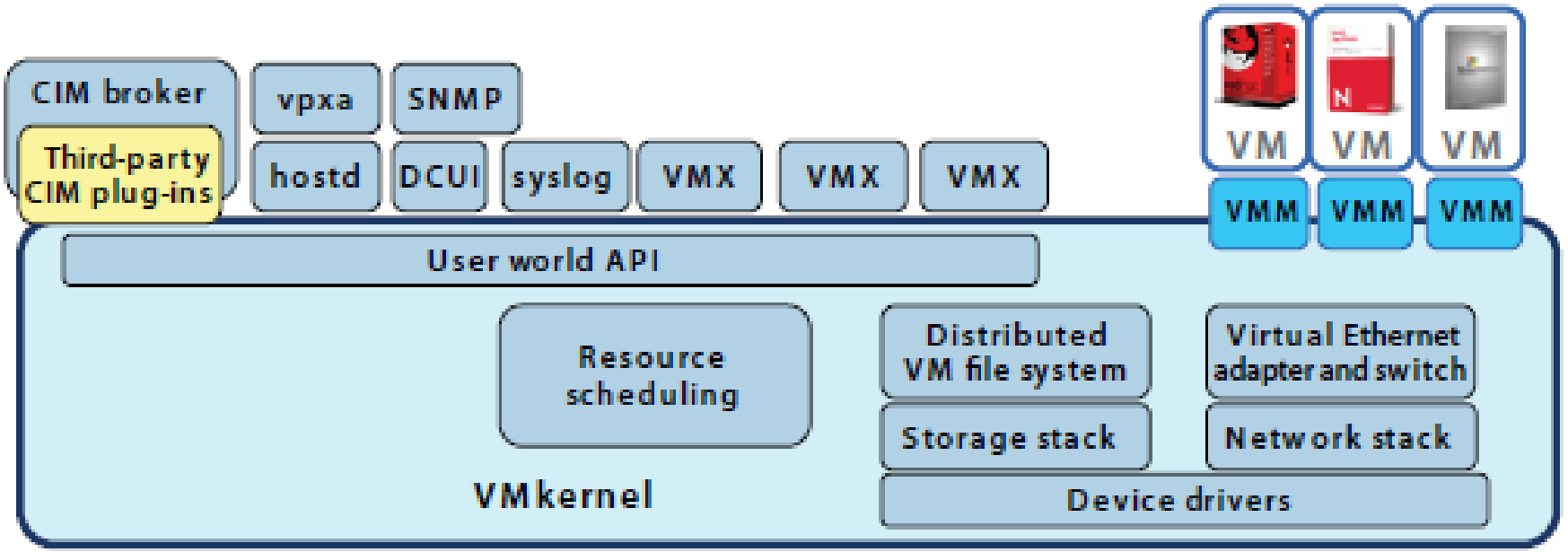




	Intel Westmere EX	IBM z14	IBM POWER8
Size	513 mm^2	696 mm^2	650 mm^2
Transistors	2.6 billion	6.1 billion	4.2 billion
Cores	4 / 6 / 8 / 10	10	12
Threads per Core	2	2	8
Maximum Frequency	3.46 GHz	5.2 GHz	4.35 GHz
L3 Cache	24 MB SRAM	128 MB eDRAM	256 MB eDRAM
Scalability	8 Sockets	24 Sockets	8 Sockets

Virtualisation pour Intel Architecture®

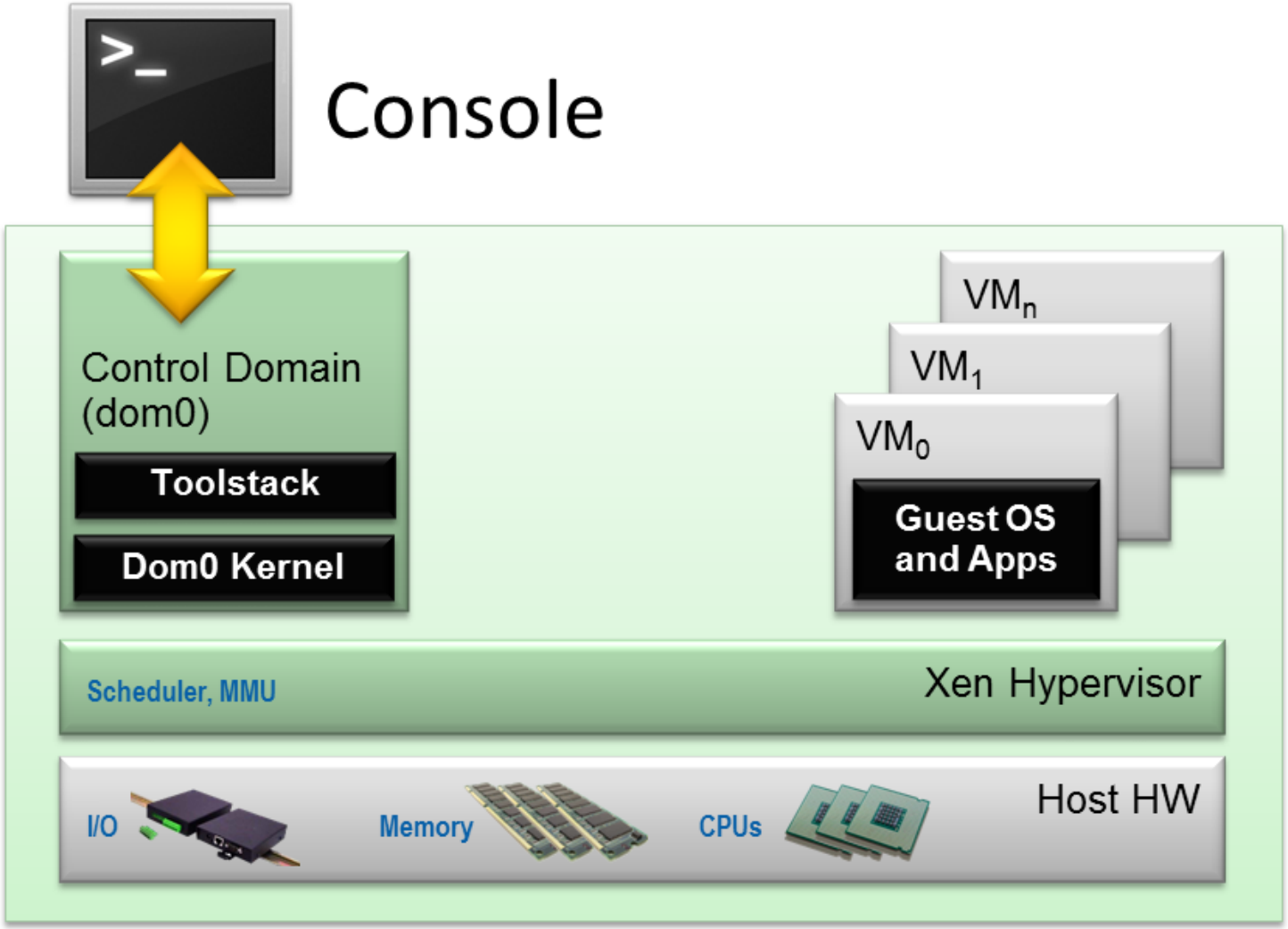
ESXi Hyperviseur de type 1



VMware vSphere 5.0

	Standard	Enterprise	Enterprise Plus
Entitlements per CPU license • vRAM Entitlement • vCPU/VM	**32 GB** **8 way**	**64 GB** **8 way**	**96 GB** **32 way**
Features • Hypervisor	✓	✓	✓
• High Availability • Data Recovery • vMotion	✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓
• Virtual Serial Port Concentrator • Hot Add • vShield Zones • Fault Tolerance		✓ ✓ ✓ ✓	✓ ✓ ✓ ✓
• Storage APIs for Array Integration • Storage vMotion • Distributed Resource Scheduler & Distributed Power Management		✓ ✓ ✓	✓ ✓ ✓
• Distributed Switch • I/O Controls (Network and Storage) • Host Profiles			✓ ✓ ✓
• **Auto Deploy*** • **Policy-Driven Storage*** • **Storage DRS***			✓ ✓ ✓
*New In vSphere 5.0			

Hyperviseur de type 1

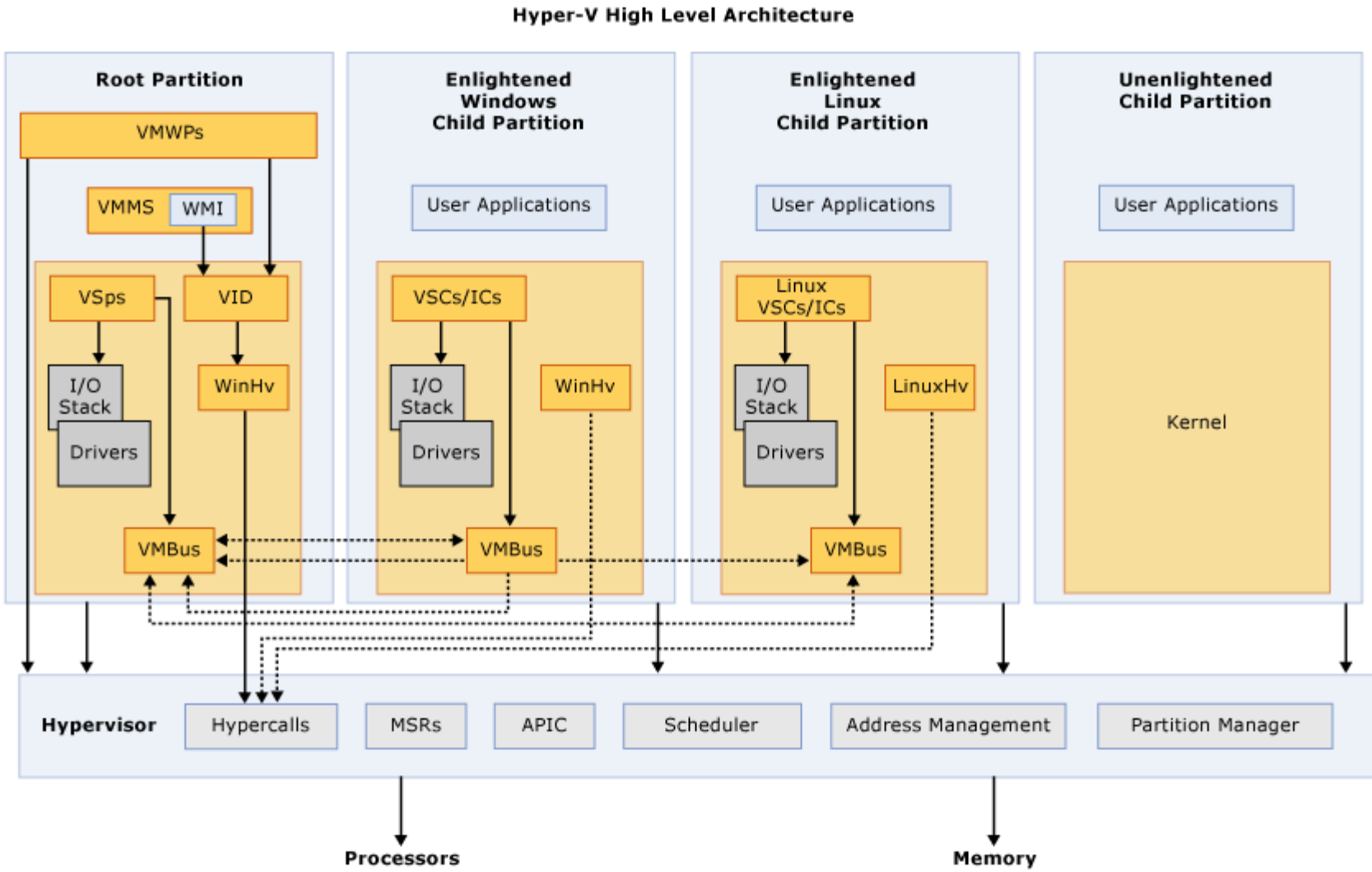


Xen limits

Server Limits	Xen 4.x
Max. physical CPUs	256
Max. dom0 virtual CPUs	256
Max. physical memory	5 TB
Max. block devices	12,000 SCSI logical units
Max. iSCSI devices	128
Max. network cards	8
Max. virtual machines per CPU core	8
Max. virtual machines per VM host	64
Max. virtual network cards	64 across all virtual machines in the system

VM Limits	Xen 4.x
Max. virtual CPUs	64
Max. memory	16 GB (32-bit), 512 GB (64-bit)
Max. virtual network devices	8
Max. virtual block devices	100 PV, 4 FV (100 with PV drivers)

Hyperviseur de type 2



Microsoft Hyperv scalability

System	Resource	Maximum number		Improvement factor
		Windows Server 2008 R2	Windows Server 2012	
Host	Logical processors on hardware	64	320	5×
	Physical memory	1 terabyte	4 terabytes	4×
	Virtual processors per host	512	1’024	2×
Virtual Machine	Virtual processors per virtual machine	4	64	16×
	Memory per virtual machine	64 GB	1 terabyte	16×
	Active virtual machines	384	1’024	2.7×
	Virtual disk size	2 terabytes	64 terabytes	32×
Cluster	Nodes	16	64	4×
	Virtual machines	1’000	4’000	4×

Virtualisation pour Power Architecture®

PowerVM Editions

PowerVM Editions offer a unified virtualization solution for all Power workloads

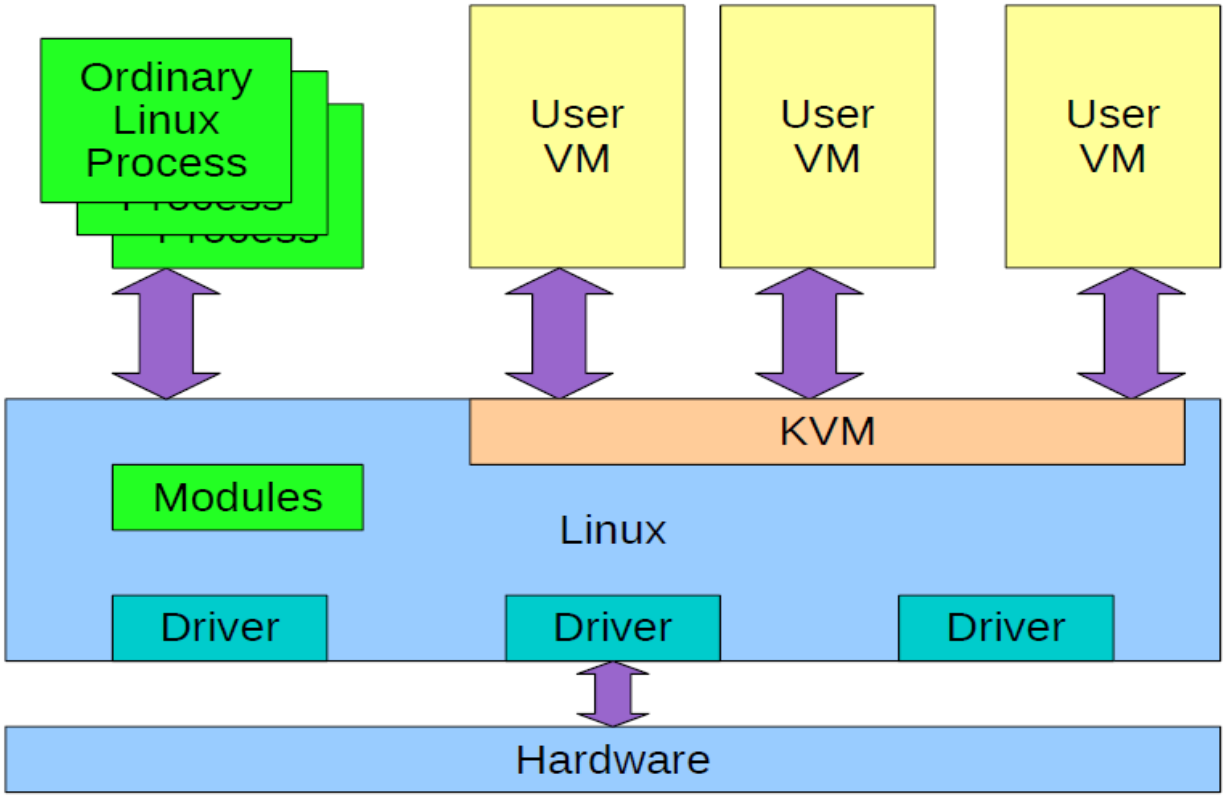
- **PowerVM Express Edition**
  - *Evaluations, pilots, PoCs*
  - *Single-server projects*
- **PowerVM Standard Edition**
  - *Production deployments*
  - *Server consolidation*
- **PowerVM Enterprise Edition**
  - *Multi-server deployments*
  - *Cloud infrastructure*

PowerVM Editions	**Express**	**Standard**	**Enterprise**
Concurrent VMs	**2 per server**	**10 per core (up to 1000)**	**10 per core (up to 1000)**
Virtual I/O Server	✓	✓ ✓	✓ ✓
NPIV	✓	✓	✓
Suspend/Resume		✓	✓
Shared Processor Pools		✓	✓
Shared Storage Pools		✓	✓
Thin Provisioning		✓	✓
Live Partition Mobility			✓
Active Memory Sharing			✓

Virtualisation avec Linux

Linux et KVM

- Hyperviseur de type 2
- Peut être utilisé sur différents processeurs (Intel, AMD, Power, zSeries …)
- Accueille les OS supportant l'architecture du processeur



IBM zEnterprise System

Chaque ordinateur répond à une ARCHITECTURE

Les publications “*z/Architecture Principles of Operation*”

https://www-05.ibm.com/e-business/linkweb/publications/servlet/pbi.wss

Search for publication

Publication number: SA22-7832

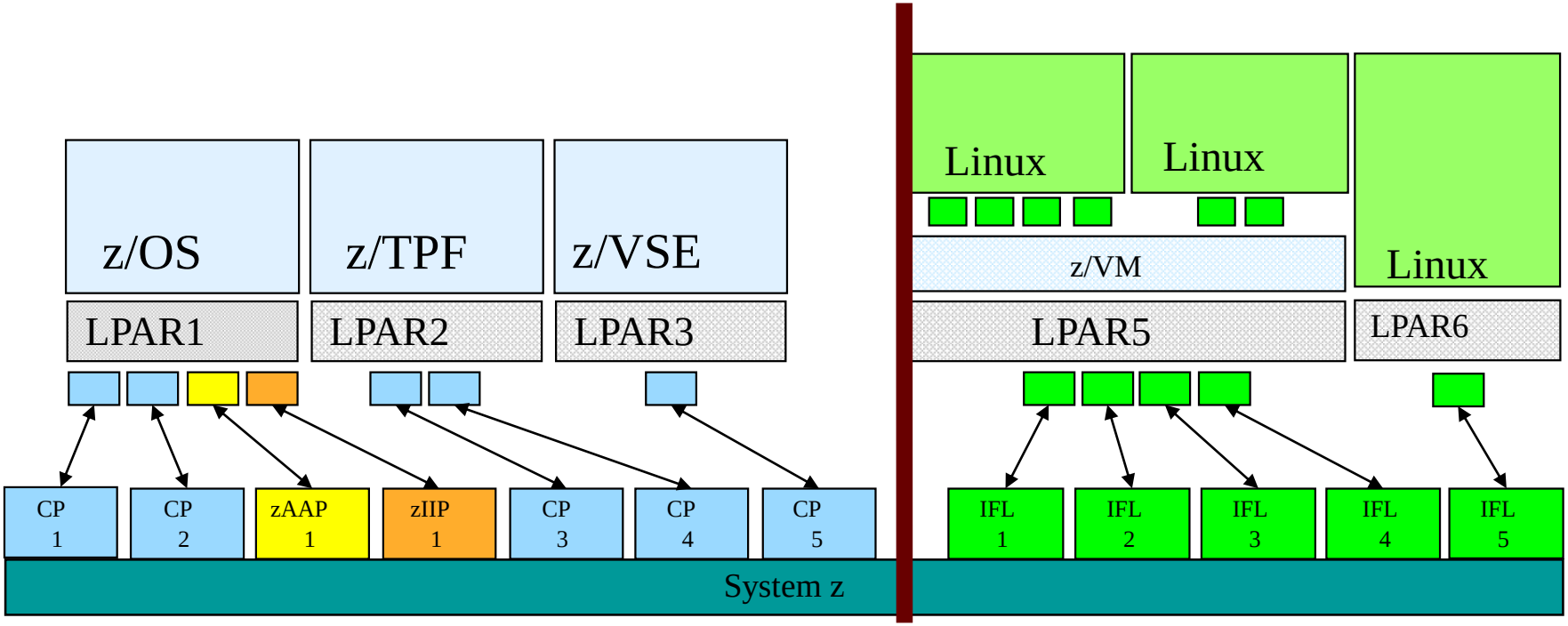
- Jeux d’instructions
- Options des Processeurs (registres, timers, interruption management)
- Disposition de la Mémoire
- Comment les E/S sont faites

Vocabulaire du System z

PC, Unix	System z
Memory	Storage
Disk	DASD (Direct Access Storage Device) – Minidisk, vdisk (z/VM)
Processor	Processor, engine, PU (Processing Unit), IOP (I/O Processor), CPU (Central Processing Unit), **CP** (Central Processor), SAP (System Assist Processor) Specialized processor **IFL** (Integrated Facility for Linux) – for Linux workload **zIIP** (z Integrated Information Processor) – Java processes
Computer	CEC (Central Electronic Complex), server
Boot	IPL (Initial Program Load)

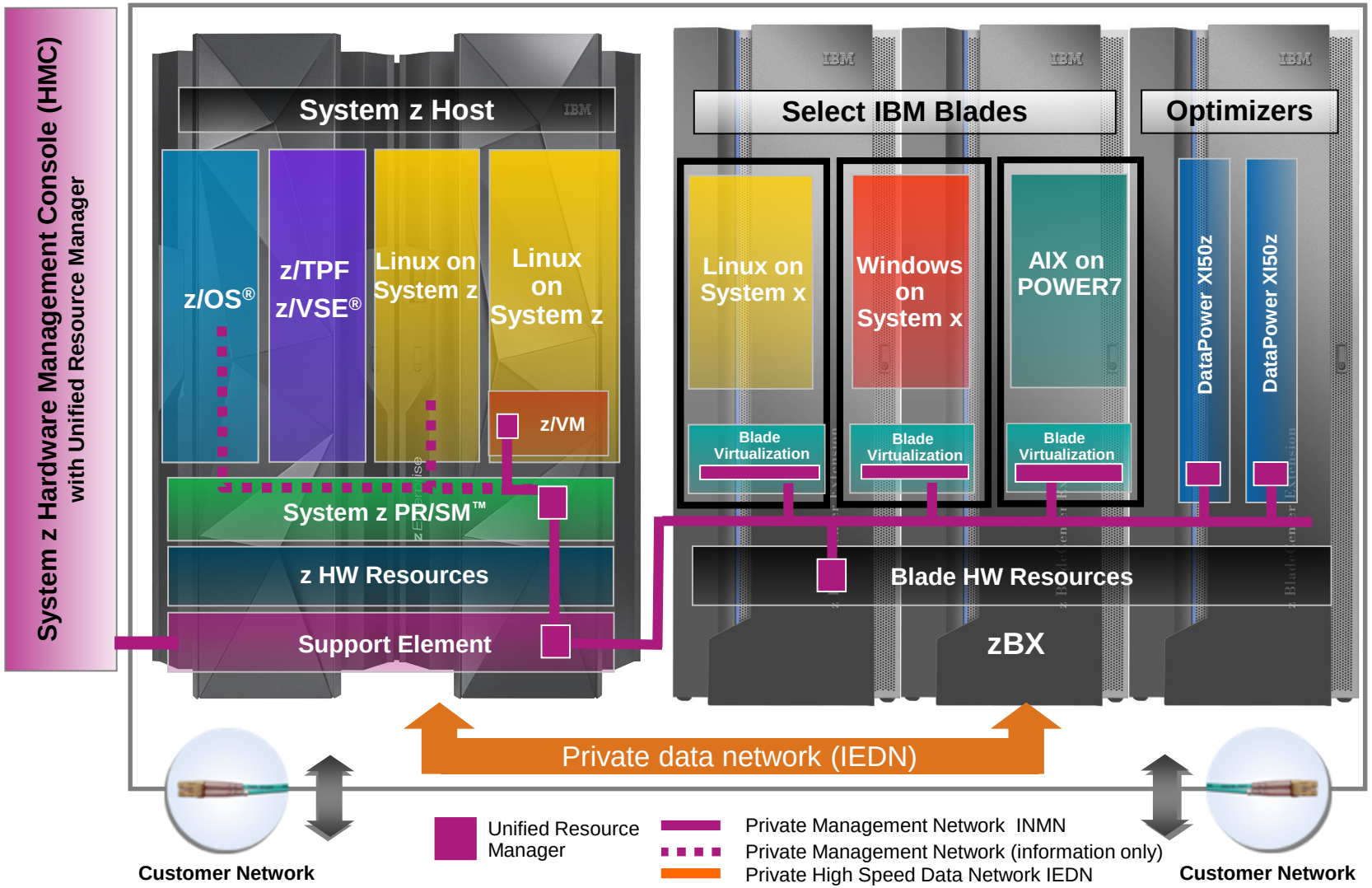
Integrated Facility for Linux

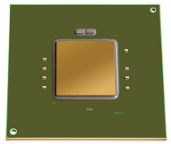
- Processeur spécialisé pour LINUX
  - z/VM & Linux for System z savent les exploiter mais pas z/OS
  - Un seul modèle de processeur à vitesse d'horloge maximum
- Les processeurs IFL ne rentrent pas dans le calcul des couts des redevances pour les logiciels z/OS, z/VSE, z/TPF car ils ne peuvent pas être exploités par ces Operating Systems.



IBM zEnterprise System

zEnterprise System (hybride)

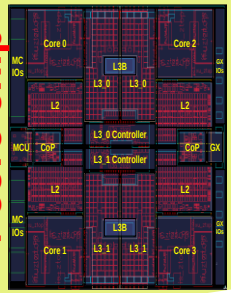




Processor roadmap

45 nm

z196/z114
9/2010



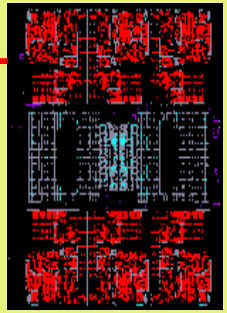
4-core chip

1.4 B transistors

- Top Tier Single Thread Performance, System Capacity
- Accelerator Integration
- Out of Order Execution
- Water Cooling
- PCIe I/O Fabric
- RAIM
- Enhanced Energy Management

32 nm

zEC12/zBC12
8/2012



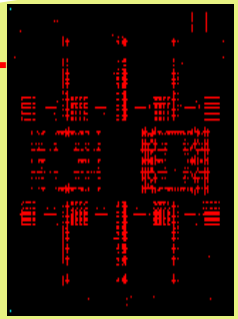
6-core chip

2.75 B Transistors

- Leadership Single Thread, Enhanced Throughput
- Improved out-of-order
- Transactional Memory
- Dynamic Optimization
- 2 GB page support
- Step Function in System Capacity

22 nm

z13/z13s
1/2015



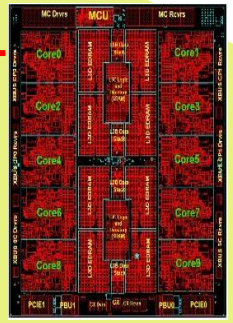
8-core chip

3.99 B Transistors

- Leadership System Capacity and Performance
- Modularity & Scalability
- SMT (2-way)
- Double instruction bandwidth
- SIMD
- Business Analytics Optimized

14 nm

z14
7/2017



10-core chip

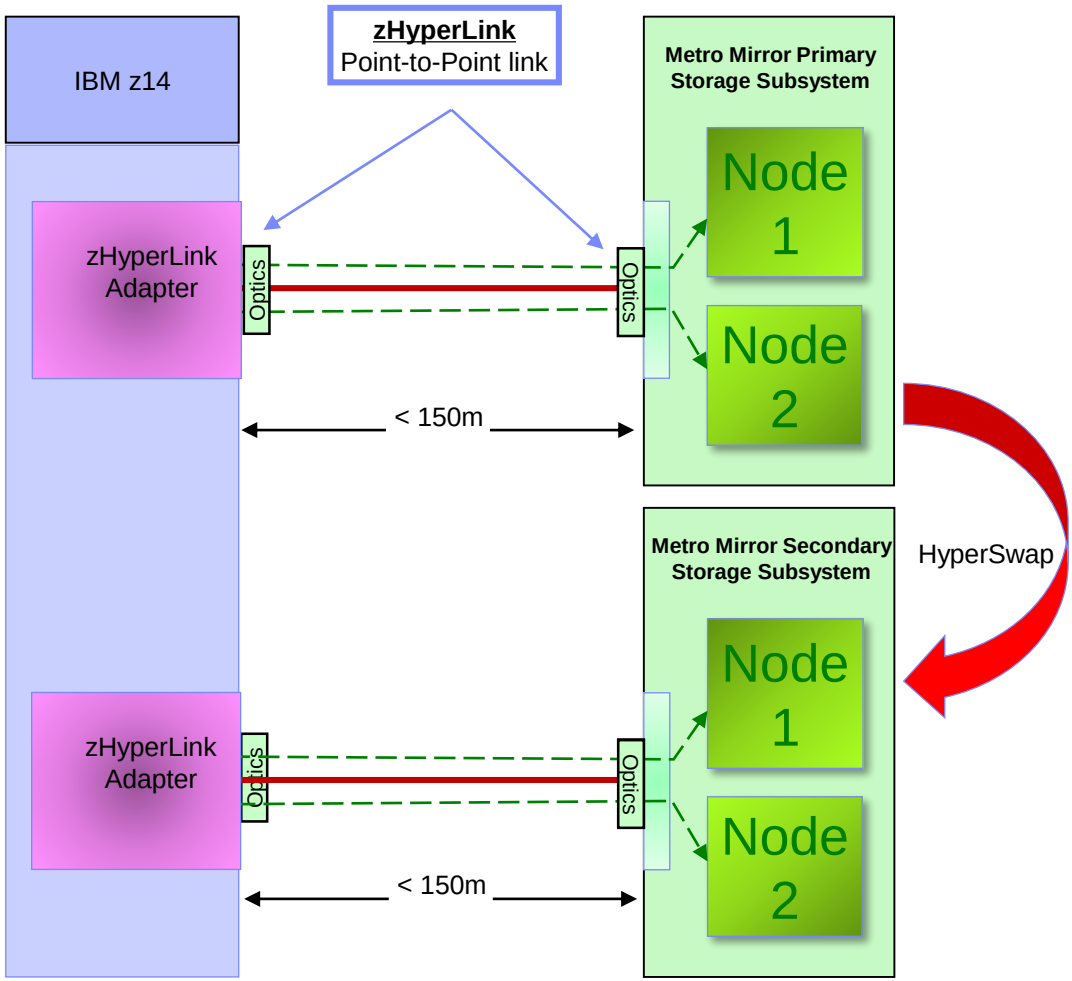
6.1 B Transistors

- Leadership System Capacity and Performance
- Improved data compression
- Enhanced hardware encryption
- Enhanced SMT (2-way)
- Enhanced SIMD
- Pause-less Garbage Collection for Java

Continuous Availability - IBM zHyperLink+ zHyperWrite

- **zHyperLink™ are point-to point-connections**
- **A maximum distance of 150m**
- **160.000 IO operations / sec**
- **8 GBytes/sec**

- **zHyperWrite™ based replication solution allows zHyperLink™ replicated writes to complete in the same time as non-replicated data**



16 zHyperLink Ports supported on each Storage Subsystem

System z I/O Subsystem Internal Bus Interconnect Speeds

PCIe
z14, z13
zEC12(8GBps)

16 GBps

InfiniBand
z10/z196/z114
zEC12

6 GBps

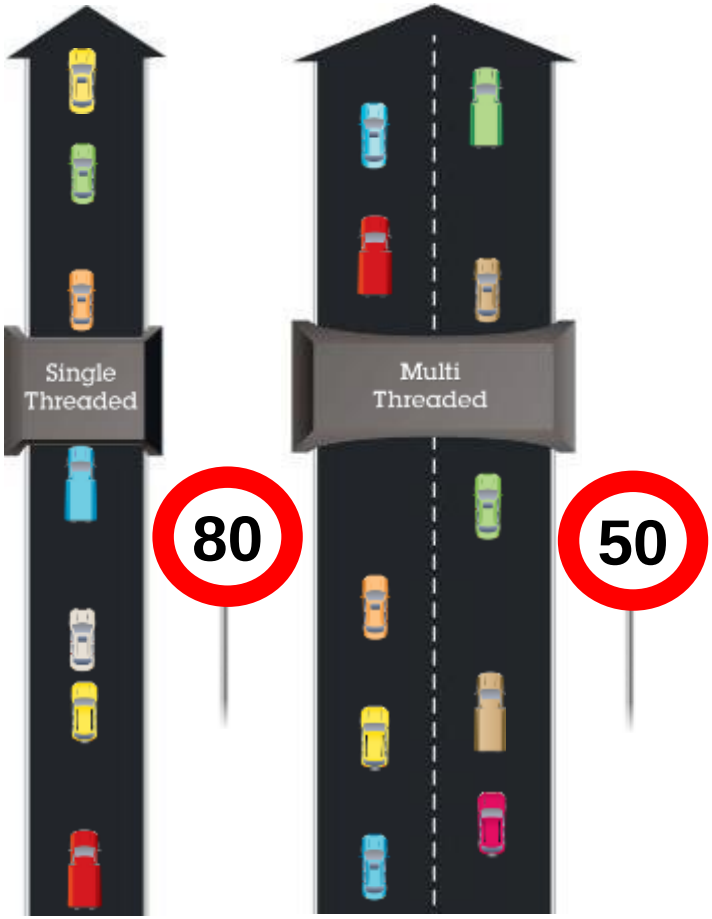
STI
z9

STI: Self-Timed Interconnect

2.7 GBps

Simultaneous Multithreading (SMT)

- **Allows instructions from one or two threads to execute on a SAP, zIIP or IFL processor core.**
- **SMT helps to address memory latency, resulting in an overall capacity* (throughput) improvement per core**
- **Each thread runs slower than a non-SMT core, but the combined 'threads' throughput is higher.**
- **The overall throughput benefit depends on the workload**
- **SMT exploitation**
  - **z/VM V6.3 + PTFs for IFLs**
  - **z/OS V2.1 + PTFs in an LPAR for zIIPs**
- **SMT can be turned on or off on an LPAR by LPAR basis by operating system parameters.**
- **z/OS can also do this dynamically with operator commands.**



*Which approach is designed for the highest volume** of traffic? Which road is faster?*

Two lanes at 50 carry 25% more volume if traffic density per lane is equal

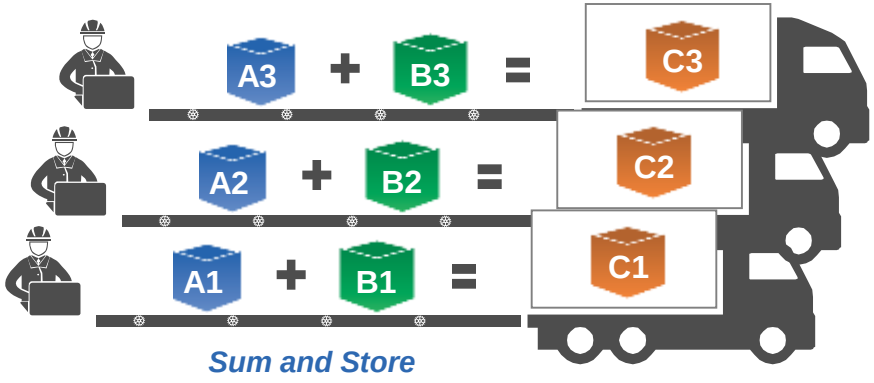
SIMD (Single Instruction Multiple Data) processing

Increased parallelism to enable analytics processing

- Smaller amount of code helps improve execution efficiency
- Process elements in parallel enabling more iterations
- Supports analytics, compression, cryptography, video/imaging processing

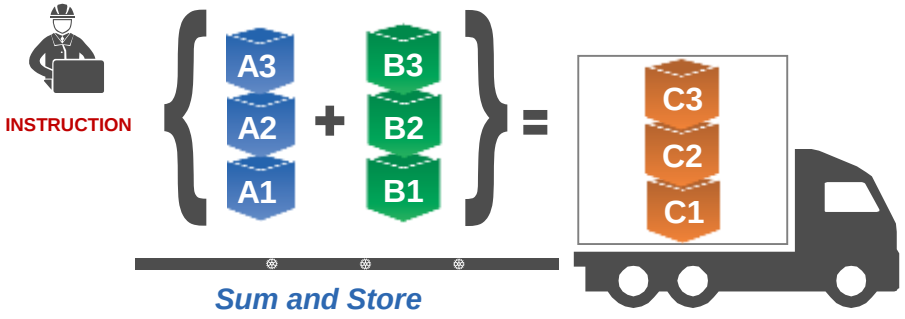
Value
- ✓ Enable new applications
- ✓ Offload CPU
- ✓ Simplify coding

Scalar
SINGLE INSTRUCTION, SINGLE DATA



Instruction is performed for *every data element*

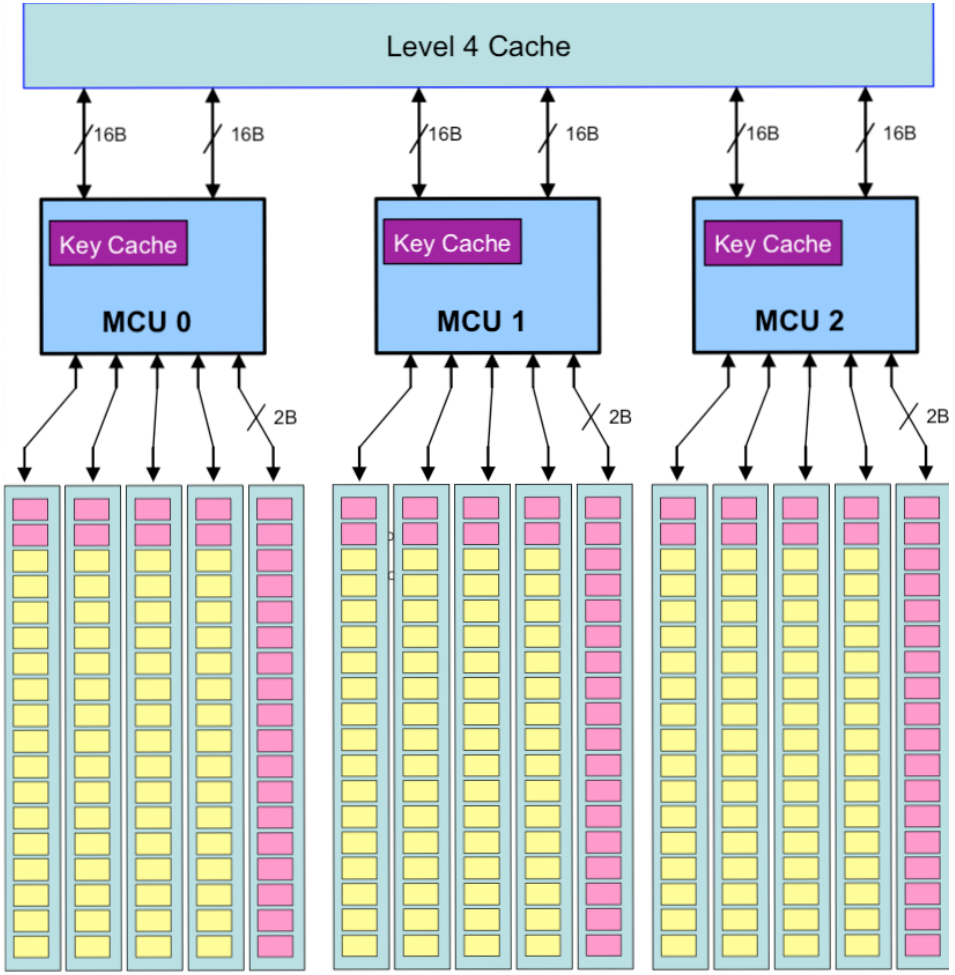
SIMD
SINGLE INSTRUCTION, MULTIPLE DATA



Perform instructions on *every element at once*

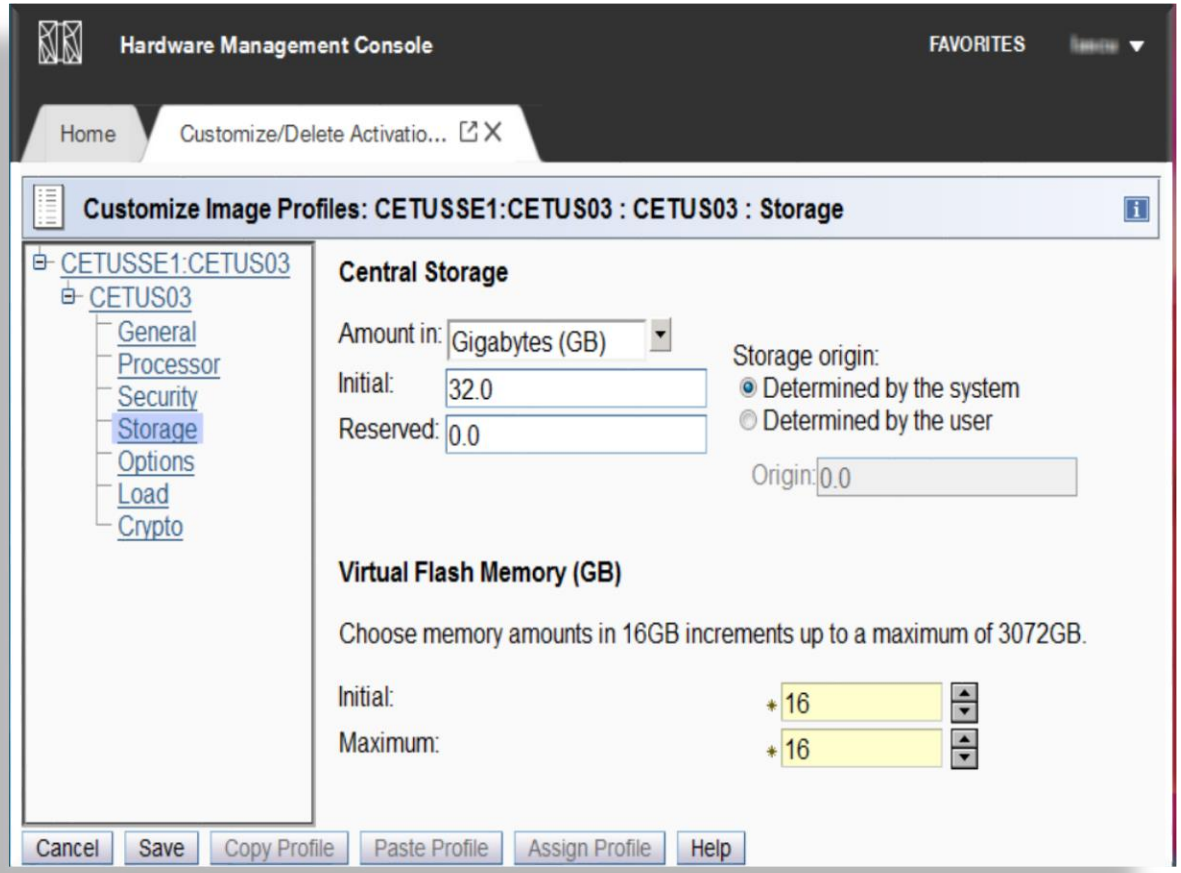
RAIM (RAID in memory)

- Issues with larger memory
  - Increases in density
  - Cosmic rays
- Improves availability
- No performance penalty



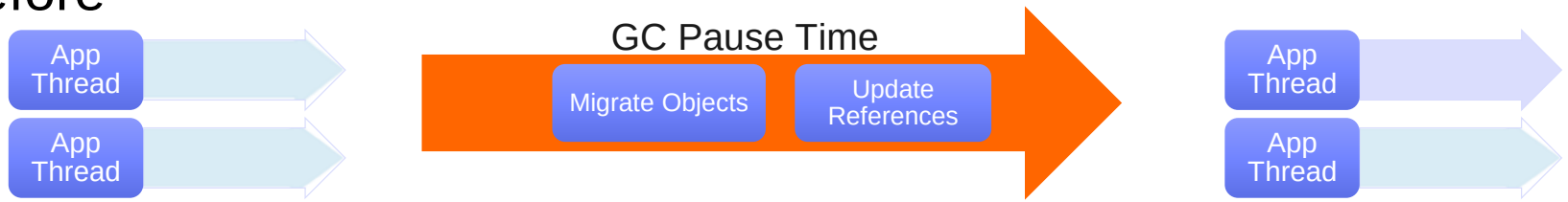
Virtual Flash Memory

- Replace the Flash Express card (4 features 1.5, 3.0, 4.5 6.0 TB)
- Now part of the zSeries memory (Protected by RAIM and ECC)
- 10% Performance improvment , less READ/WRITE latency (/1000)
- Does not use the PCIe interface

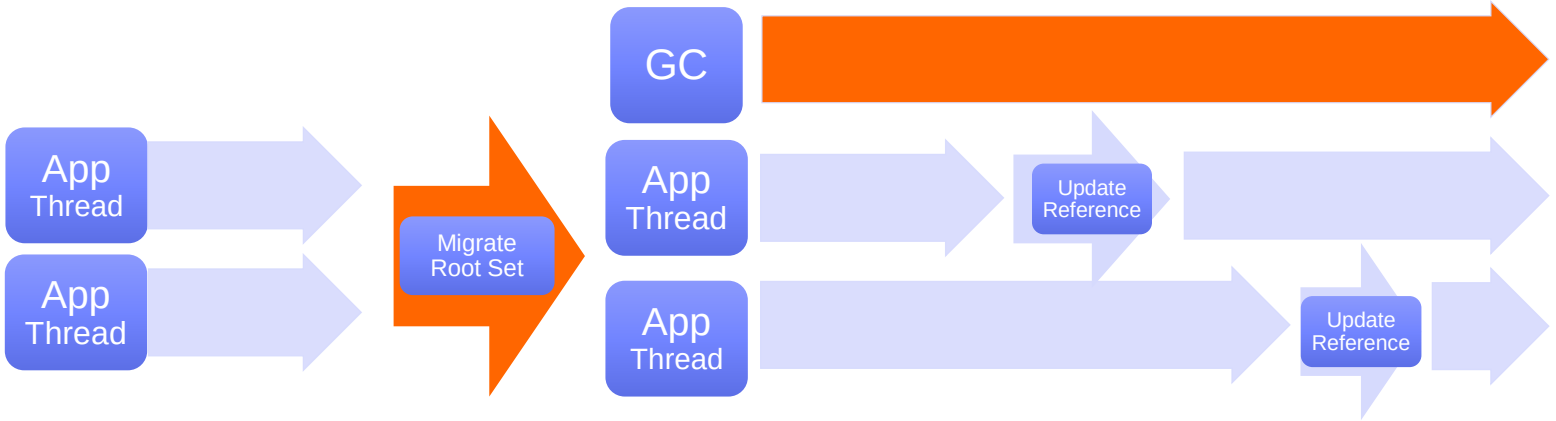


JAVA Garbage Collection – Enhancements

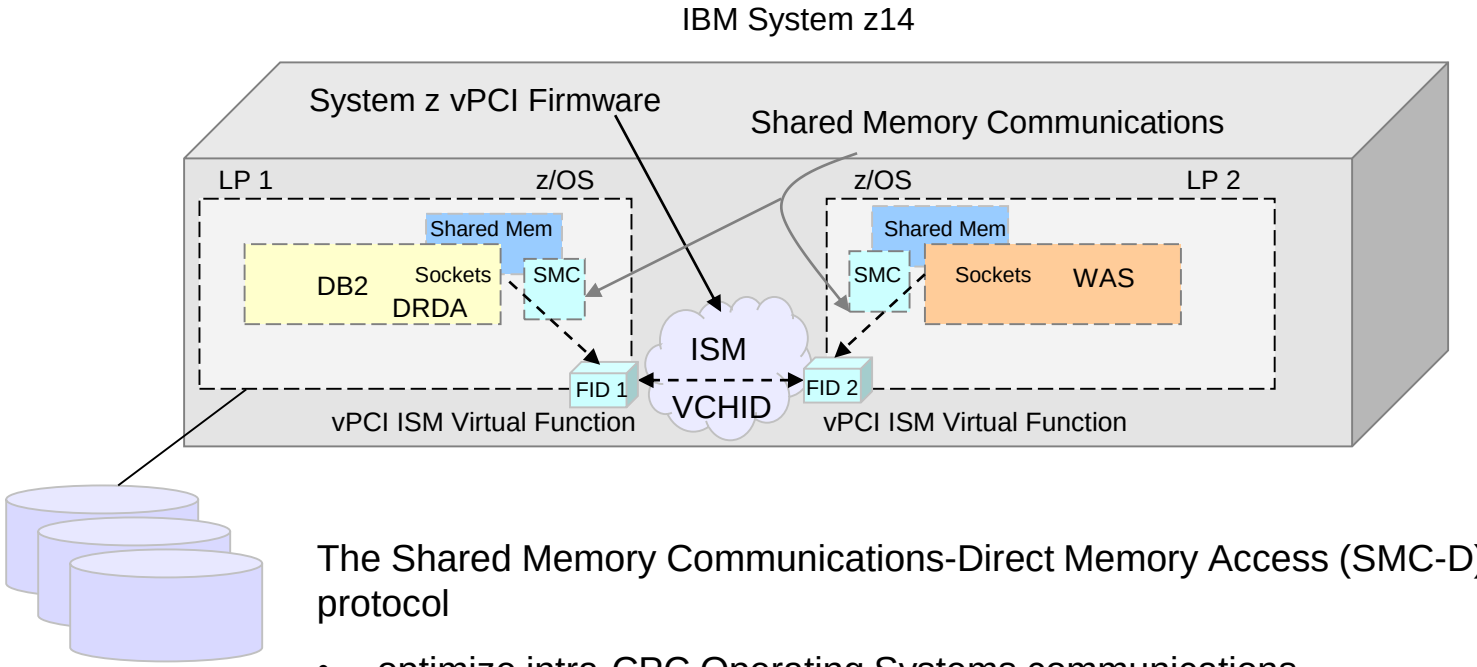
Before



Using a z14



SMC – Shared Memory Communications

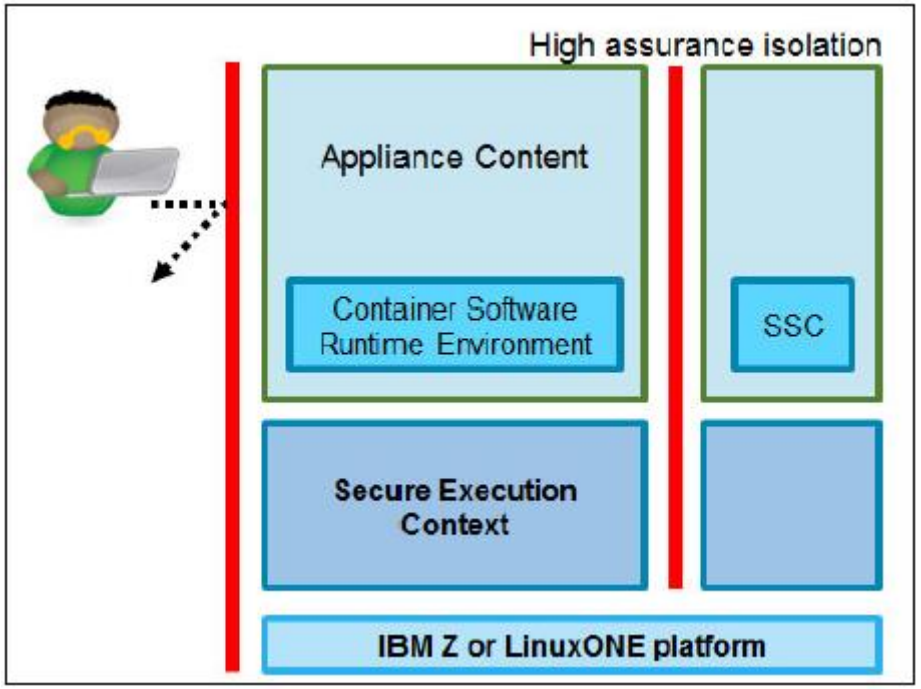


The Shared Memory Communications-Direct Memory Access (SMC-D) protocol

- optimize intra-CPC Operating Systems communications
- transparent to socket applications
- tightly couples socket API communications / memory within the CPC
- eliminates TCP/IP processing in the data path
- ISM is a z System firmware solution (no additional hardware required).

Secure Socket Containers

- Highly secure context
- No system admin access
- Only remote API
- Memory access of system admin is disabled
- Uses encrypted disk
- Encrypted dumps
- Strong isolation between containers
- High assurance isolation

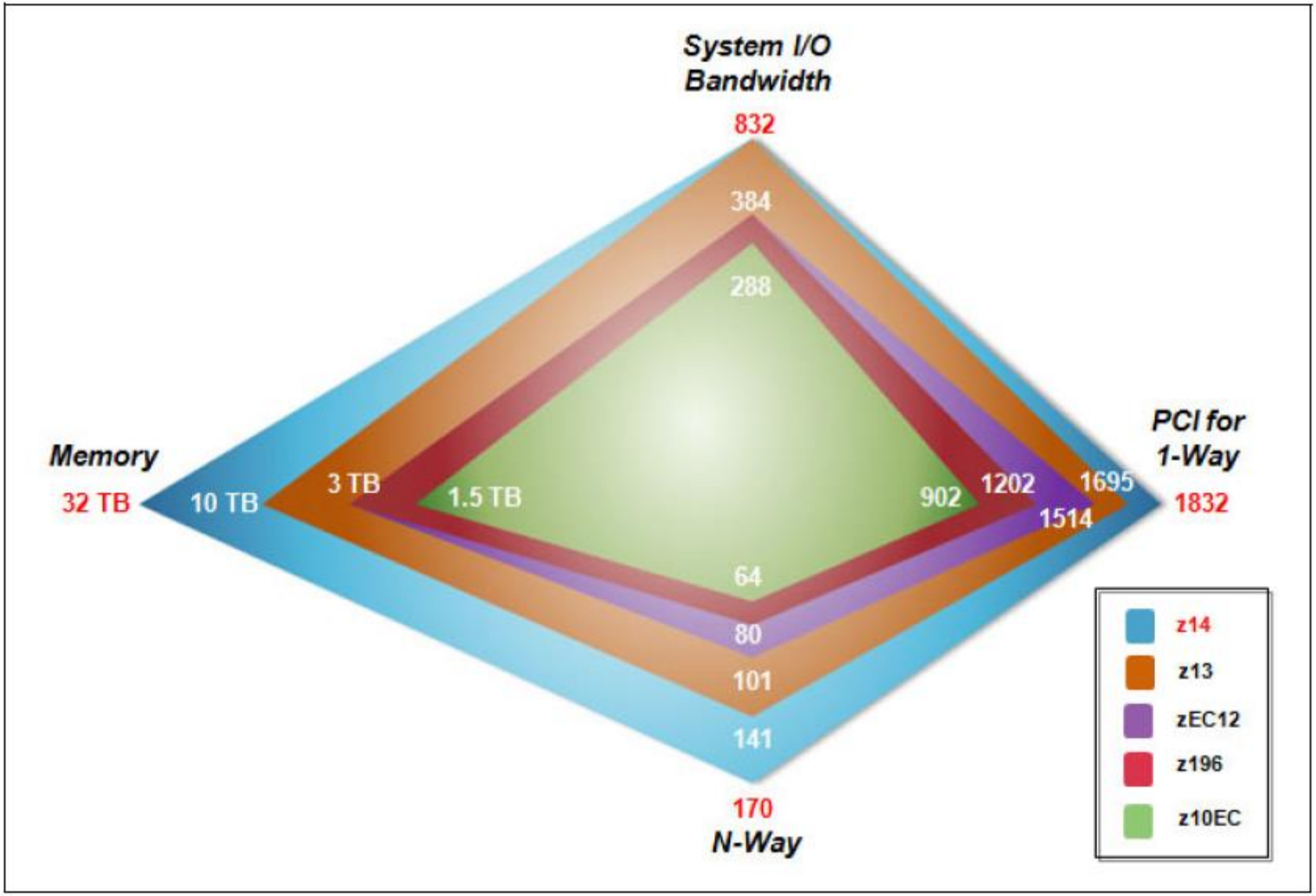


FICON availability

- **Forward Error Correction Code (FEC)**
  - Allows FICON channels to operate at higher speeds, over longer distances, with reduced power and higher throughout while retaining traditional RAS levels
  - Used for controlling errors in data transmission over unreliable or noisy communication channels
    - Sender encodes messages in a redundant way by using error-correcting code (ECC)
    - Allows the receiver to detect a limited number of errors that might occur anywhere in the message and often corrects these errors without retransmission

- **FICON Dynamic Routing (FIDR)**
  - Enables exploitation of SAN dynamic routing polices in the fabric to lower cost and improve performance for supporting I/O devices
    - Share ISLs for FICON traffic and FCP Metro Mirror traffic
    - Better utilization of available ISL bandwidth
    - Simplified Management of ISLs

System z: Design Comparison



IBM System z Generations

	zEC12	z13	z14
Maximum cores or IFLs	101	141	170
#LPARs	60	85	85
zIIPs	Optional	Optional	Optional
Standard SAPs	16	24	23
Minimum Memory	32 GB	64 GB	256 GB
Max Orderable Memory	3 TB	10 TB	32 TB
Co-processor compression	Standard	Standard	Standard
Shared Memory Communcations - Direct Access Memory (SMC-D)	Not supported	Optional	Optional
SIMD	Not supported	Standard	Standard
Simultaneous Multi Threading	Not supported	Standard	Standard
Java Garbage Collection	Not supported	Not supported	Standard

IBM System z Generations

	zEC12	z13	z14
I/O			
FICON Express	8S, 8	8S, 16S	16S+, 16S
zEDC Express	Optional	Optional	Optional
Flash Express	Optional	Optional	Not supported
FICON Dynamic Routing	Not supported	Standard	Standard
IBM zHyperLinks	Not supported	Not supported	Optional
IBM Virtual Flash Memory	Not supported	Not supported	Optional
HiperSockets	Standard	Standard	Standard
Networking			
10 GbE RoCE Express	Optional	Optional	RoCE Express2
Shared Memory Communications - RDMA (SMC-R)	Optional	Optional	Optional

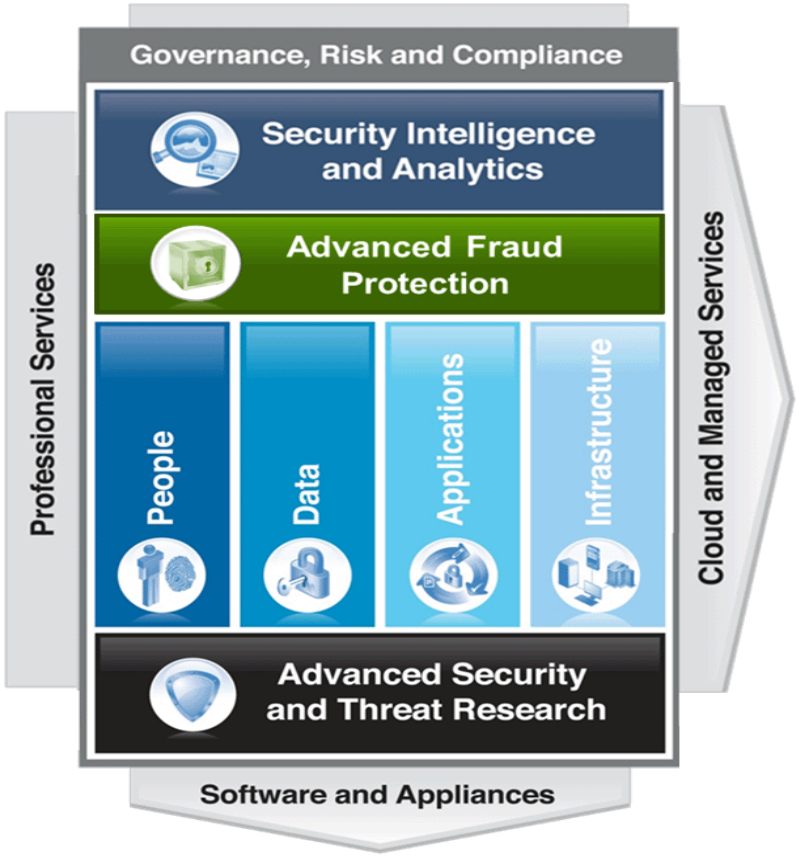
IBM System z Generations

	zEC12	z13	z14
Security			
Crypto Express	Crypto Express4S	Crypto Express5S	Crypto Express6S
EAL5+	Standard	Standard	Standard
CPACF (CP Assist for Cryptographic Functions)	Optional	Optional	Optional
EMV (Europay, Mastercard and Visa)	Standard	Standard	Standard
API			
Secure Service Container	Not supported	Standard	Standard

IBM System z Generations

	zEC12	z13	z14
Availability			
Forward Error Correction (FEC)	Not supported	Standard	Standard
IC Coupling Links	Supported	Supported	Supported
Coupling Express LR	Not supported	Optional	Optional
12x InfiniBand Coupling Links	Optional	Optional	Optional
1x Infiniband Coupling Links	Optional	Optional	Optional
ICA Short Range Coupling Links	Not supported	Optional	Optional

Security



- **ACCES & IDENTITE**
- **SECURITE DES DONNEES**
- **SECURITE DES APPLICATIONS**
- **SECURITE DES INFRASTRUCTURES**
- **CONFORMITE DE LA SECURITE**

Elements of the System z Enterprise Security Hub

CPACF

Crypto Express Cards

Tape encryption

TS1120

Disk encryption

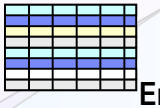
DS8000®

Secured Key Storage & Management

DKMS
TKLM

Optim™

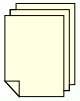
Multilevel Security



Guardium

Enterprise Fraud Solutions

System z SMF



IBM Tivoli Security Compliance Insight Manager

PKI Services

Certificate Authority

Tivoli Identity Manager

IBM Tivoli® zSecure Suite

Compliance and Audit

Extended Enterprise

DB2® Audit Management Expert

Data Privacy

Platform Infrastructure

Tivoli Federated Identity Mgr

Common Criteria

Common Criteria Ratings Support for Standards

z/OS Java SDK

Optimized for z/OS

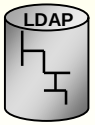
RACF®

Audit, Authorization, Authentication, and Access Control

ICSF

Services and Key Storage for Key Material

ITDS



Scalable Enterprise Directory

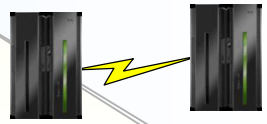
Network Authentication Service

Kerberos V5 Compliant

z/OS® System SSL

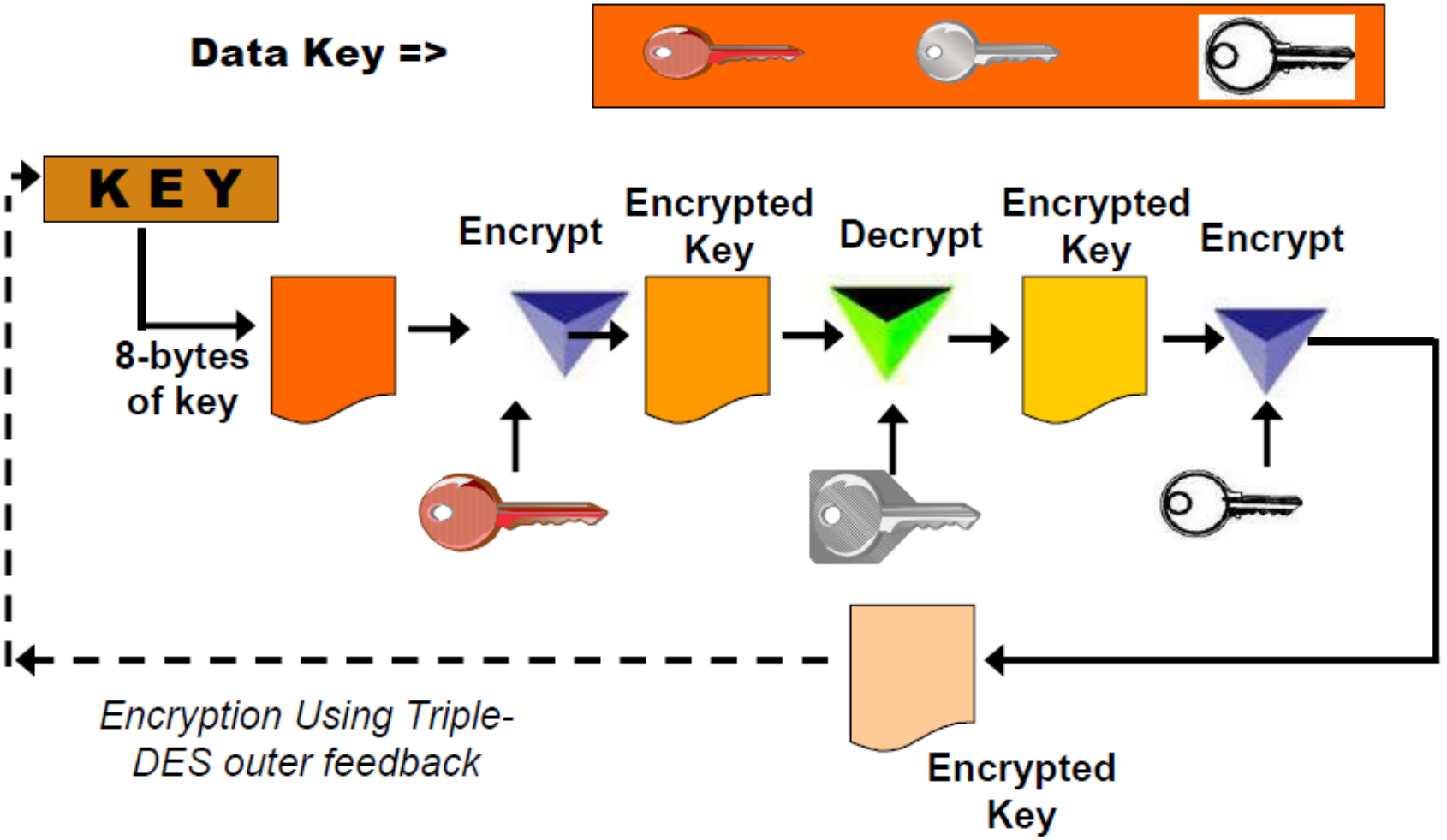
SSL/TLS suite

Communications Server



IDS, Secure Communications

Data Confidentiality – DES/TDES



Data Confidentiality - AES

- **Rijndael Algorithm**
  - Block Cipher (16-byte blocks)
  - 128-, 192, 256-bit Key Length
  - Multiple Rounds
  - Four Steps per Round (Byte Substitution, Shift Row, Mix Column, Add Round Key)

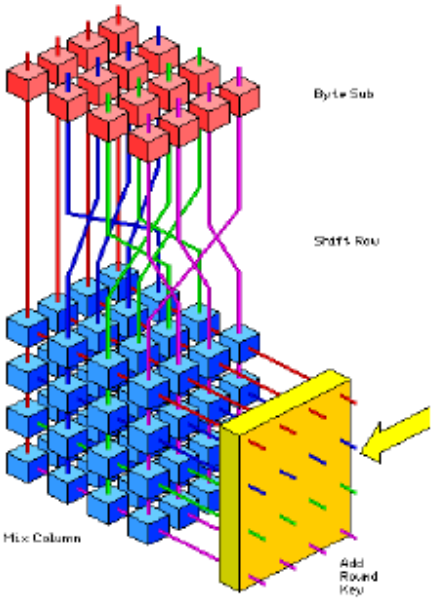


Image from http://www.esat.kuleuven.ac.be/~rijmen/rijndael/

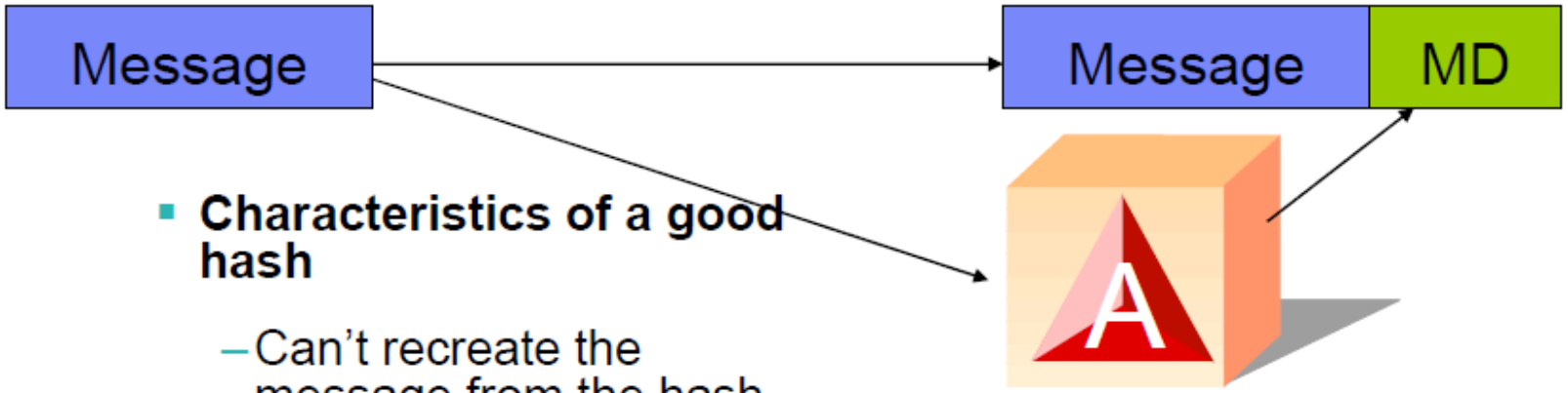
Public Key Architecture - PKA

- **Asymmetric Keys**
  - RSA, Rivest Shamir and Adleman
  - Diffie-Hellman
  - Elliptic Curve (ECC)

Encipher Key ≠ **Decipher Key**

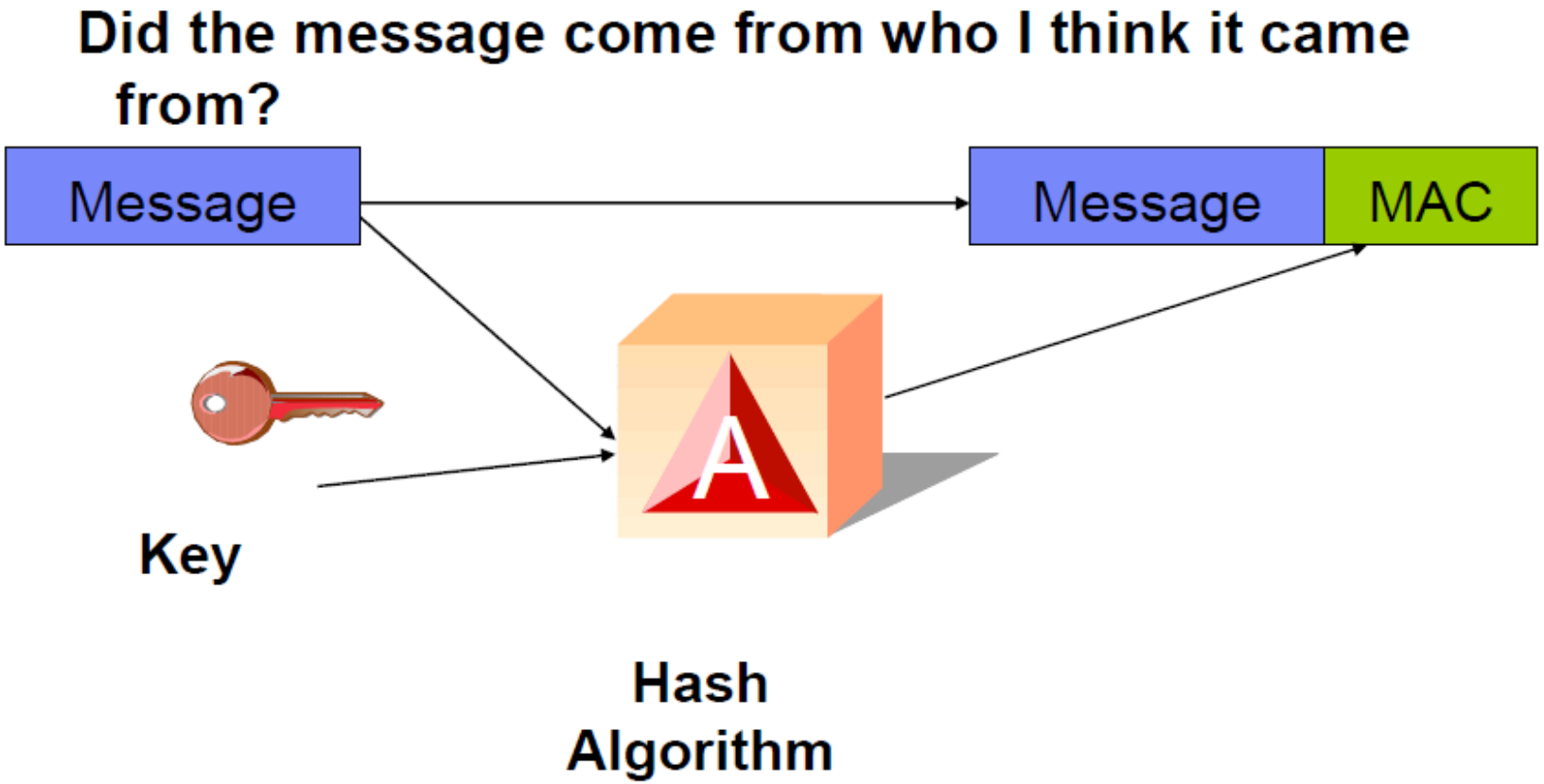
Data Integrity – Modification Detection

Has the message changed?



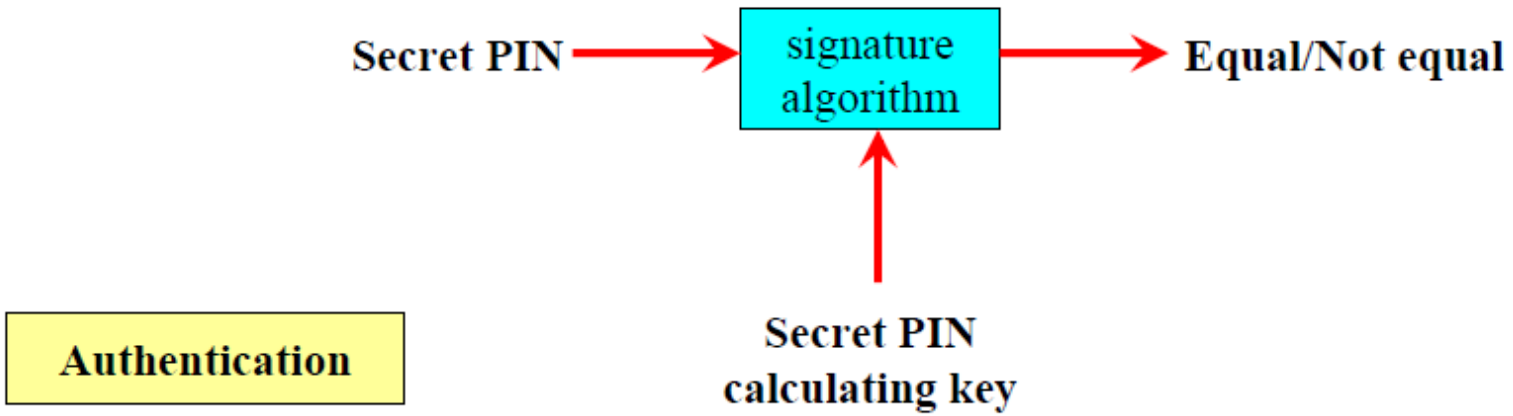
- **Characteristics of a good hash**
  - Can't recreate the message from the hash
  - Two different messages are statistically unlikely to generate the same hash value

Data Integrity – Message Authentication



Financial Services

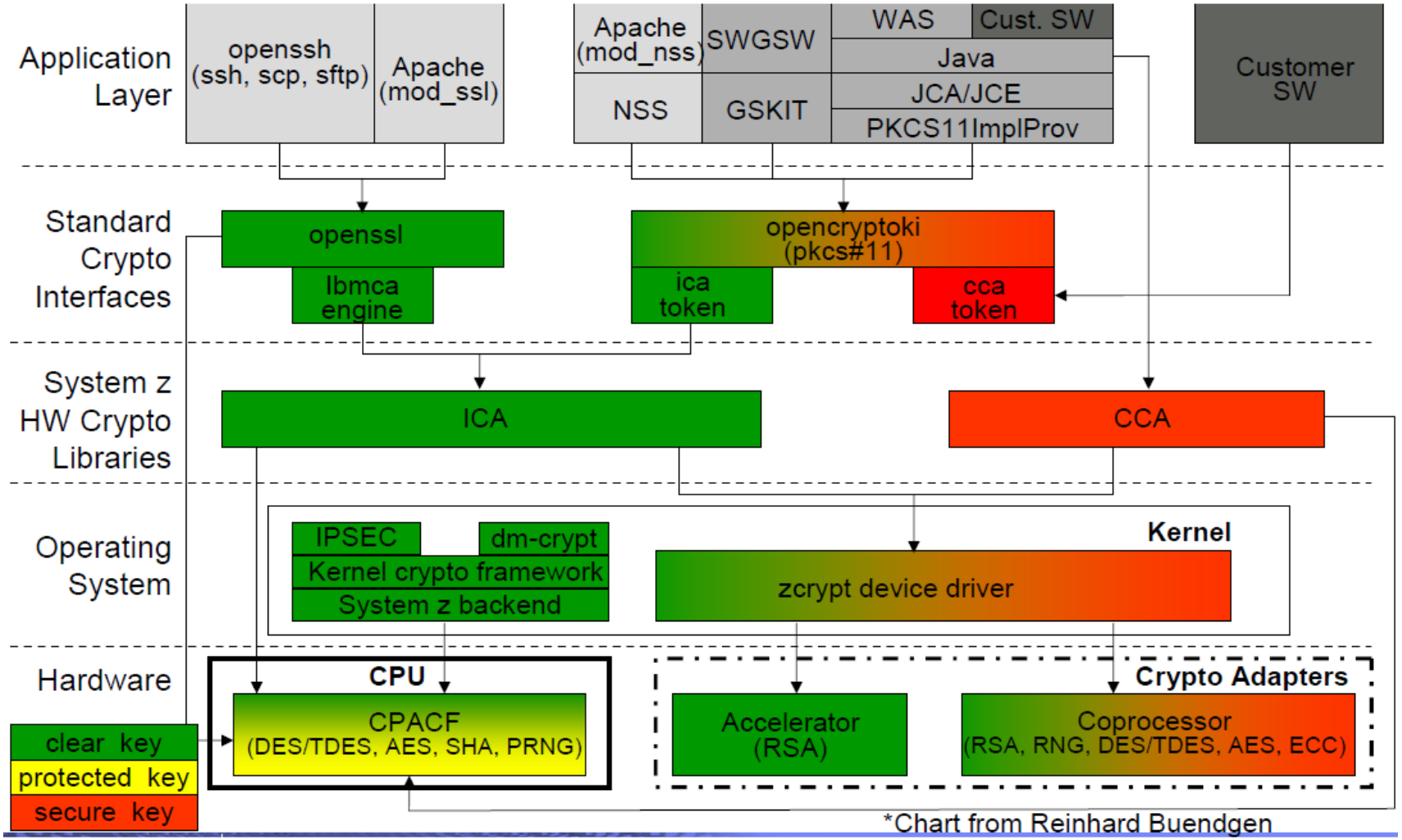
- **PIN Generation**
- **PIN Verification**
- **PIN Export/Import**



Key security

- Clear Key
  - key may be in the clear, at least briefly, somewhere in the environment
- Secure Key
  - key value does not exist in the clear outside of the HSM (secure, tamper-resistant boundary of the card)
- Protected Key
  - key value does not exist outside of physical hardware, although the hardware may not be tamper-resistant

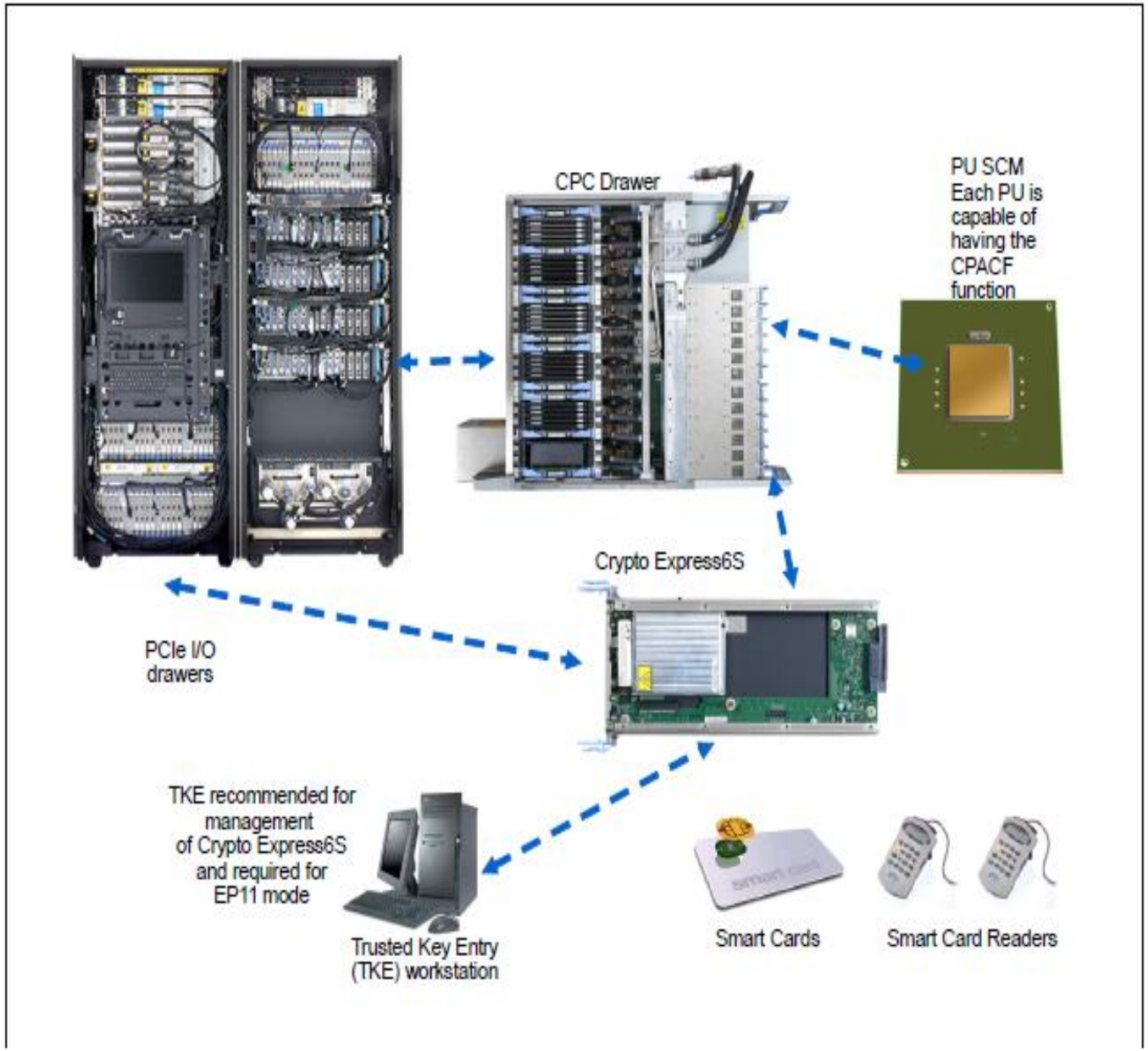
Security with Linux for zSeries



z14 and crypto

CPACF

- DES, TDES, AES-128, AES-192, AES-256, SHA-1,SHA-256, SHA-384, SHA-512, PRNG, DRNG, TRNG



Crypto Express6S

Data privacy and confidentiality

DES Data Encryption Standard 56 bits key
TDES Triple DES – DES applied 3 times (3x56 bits key)
AES Advanced Encryption Standard (128 or 192 or 256 bits key
SHA1 Secure Hash Algorithm 160-bit
SHA2 224- 256- 384- and 512-bit
RSA <= 4K-bit
ECC Elliptic Curve Coding

Key generation

PRNG Pseudo Random Number Generator
RNG 4096-bit RSA key
RNGL 8 to 8096 bytes

Message authentication code

Single or double key MAC

One PCIe adapter per feature (Initial order – two features)

FIPS 140-2 Level 4

Up to 16 features per server

- Three configuration options
- Secure IBM CCA coprocessor:
  - For secure key encrypted transactions using CCA callable services (default).
- Accelerator:
  - For public key and private key cryptographic operations that are used with Secure Sockets Layer/Transport Layer Security (SSL/TLS) acceleration.
- Secure IBM Enterprise PKCS #11 (EP11) coprocessor:
  - Implements industry standardized set of services that adhere to the PKCS#11.

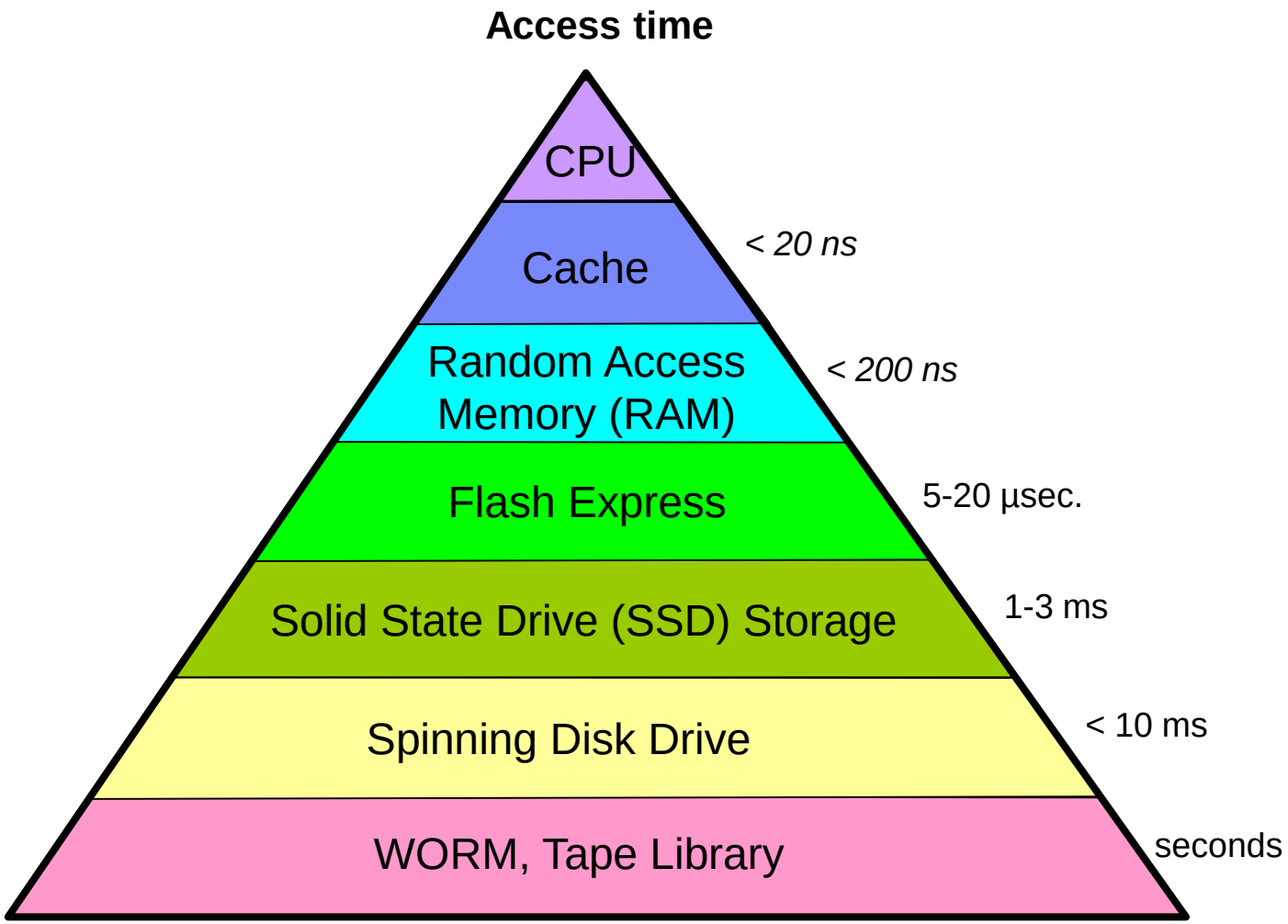
Reliability, Availability, Serviceability

Pour afficher 99,999% de disponibilité,

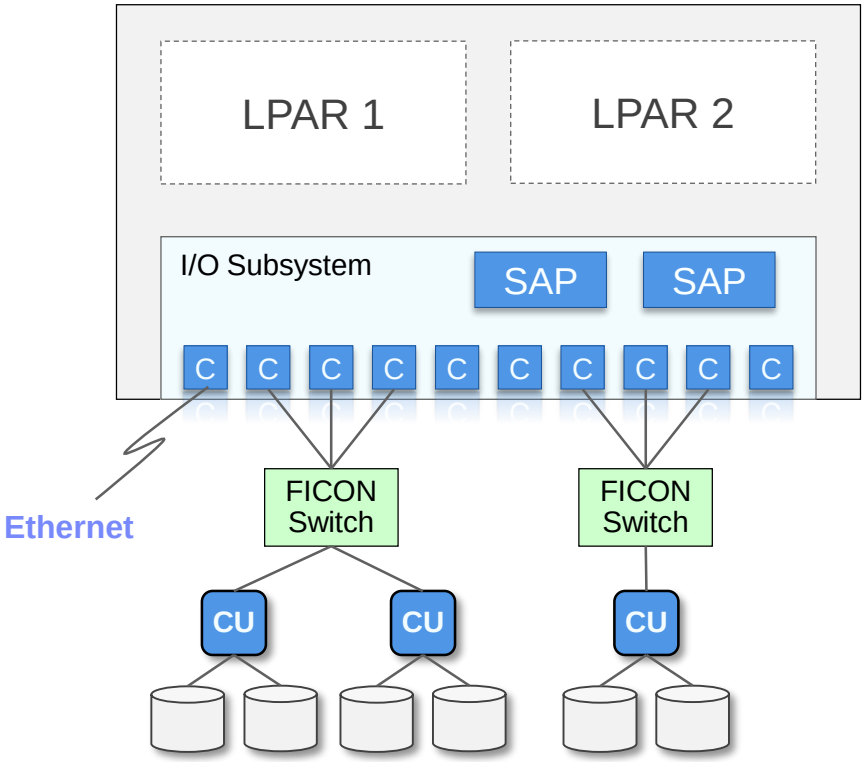
il faut avoir une très bonne disponibilité au niveau matériel et logiciel comme :

- Des alimentations redondantes
- Processeur de service et surveillance en double (SE – Support Element)
  - Génère automatiquement des appels au service de maintenance lorsqu'une erreur est détectée
  - Diagnostique à distance des éléments défectueux
  - Mise à jour du microcode sans arrêt de la machine
- I/O
  - Plusieurs processeurs gèrent les Entrées/Sorties
  - En cas de défaut, un processeur remplaçant est activé automatiquement et le défectueux est désactivé
  - Plusieurs chemins d'accès vers chaque unité de contrôle (online/offline depuis l'OS)
- Mémoire
  - RAIM (Redundant Array of Independent Memory) détecte et corrige les erreurs au niveau de la DRAM, du socket, du canal mémoire ou de la DIMM.
  - Utilise des clefs mémoire pour chaque page de 4Ko
  - Le changement, l'augmentation ou la diminution de la mémoire est fait dynamiquement
- Processeur
  - Ajout et suppression dynamique (commande dans l'OS pour mettre ON ou OFF des processeurs)
  - Contrôle des processeurs par les processeurs de service (SE - Support Element) et désactivation automatique des processeurs défectueux

Relative Access Times for different technologies



System z I/O Subsystem Offloads Work



System Assist Processor

- Manages path selection
- Filter intermediate interrupts
- I/O Priority
- Dynamic path reconnect

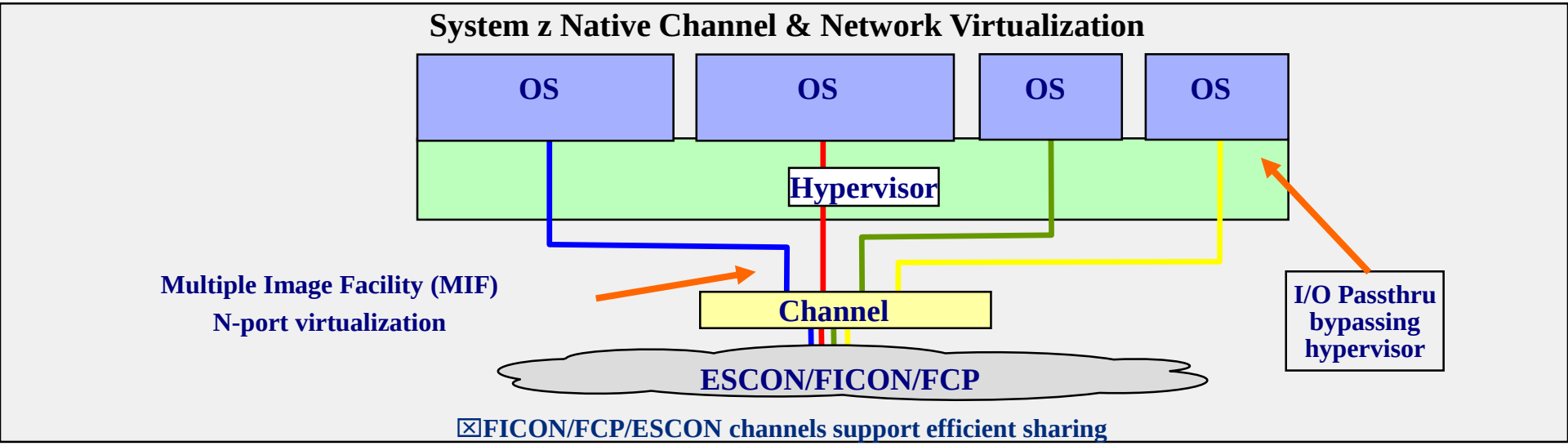
Channels

- Manages physical links
- Direct memory transfer

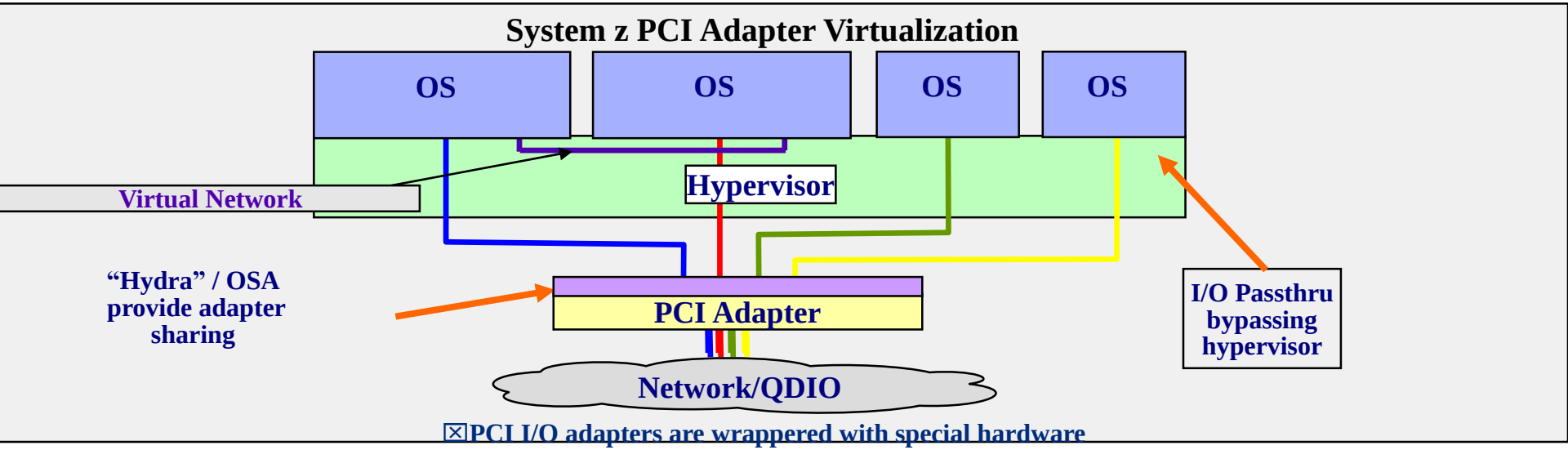
Control Units

- Manages interfaces
- Cache
- Abstracts devices
- Dynamic path reconnect

Virtualisation I/O avec le zSeries



☒FICON/FCP/ESCON channels support efficient sharing



☒PCI I/O adapters are wrappered with special hardware

Input/Output Configuration Program

Décrit les unités d'Entrées/Sorties et les chemins d'accès.

Chaque unité à une adresse entre 0000 et FFFF et est liée à un CSS (Channel SubSet - de 1 à 4 par CPU)

Un exemple d'IOCDS

```
ID       MSG1='Z114_M05 ',MSG2='',SYSTEM=(2818,1),
               TOK=('VM-
   TOKEN',F1F061F0F161F1F2F1F17AF3F57AF1F640404040*
               ,00000000,'10/01/12','11:35:16','','')
      RESOURCE PARTITION=(CSS(0),(VM04,1))
*IOCP ***********************************
*IOCP *   Définition des chpid's        *
*IOCP ***********************************
*
*IOCP * CHPIDs OSA Express 3 avec 2 ports
*
* Le Chpid 00 est un OSA ICC sur port 0
CHP00  CHPID   PATH=(CSS(0),00),TYPE=OSC,SHARED,PCHID=200
      CNTLUNIT CUNUMBR=1300,PATH=00,UNIT=OSA
      IODEVICE ADDRESS=(1300,254),MODEL=X,CUNUMBR=(1300),UNIT=3270
* carte OSA en mode OSD (QDIO)
CHP01  CHPID   PATH=(CSS(0),01),TYPE=OSD,SHARED,PCHID=210
      CNTLUNIT CUNUMBR=1400,PATH=01,UNIT=OSA
      IODEVICE ADDRESS=(1400,32),CUNUMBR=1400,UNIT=OSA
      IODEVICE ADDRESS=(14FE,1),CUNUMBR=1400,UNIT=OSAD
* carte OSA en mode OSD (QDIO)
CHP02  CHPID   PATH=(CSS(0),02),TYPE=OSD,SHARED,PCHID=220
      CNTLUNIT CUNUMBR=1100,PATH=02,UNIT=OSA
      IODEVICE ADDRESS=(1100,32),CUNUMBR=1100,UNIT=OSA
      IODEVICE ADDRESS=(11FE,1),CUNUMBR=1100,UNIT=OSAD
*IOCP *
*IOCP * CHPIDs ESCON 15 ports
*IOCP * Escon convertisseur Optica avec sortie parallele 3174
*IOCP *
CHP04  CHPID   PATH=(CSS(0),04),TYPE=CVC,PART=(VM04),PCHID=260
      CNTLUNIT CUNUMBR=0820,PATH=04,SHARED=Y,UNIT=3174,
               UNITADD=((20,32))
      IODEVICE ADDRESS=(820,32),CUNUMBR=0820,UNIT=3278
*IOCP *
*IOCP * Escon vers 3490s
*IOCP *
CHP05  CHPID   PATH=(CSS(0),05),TYPE=CNC,PART=(VM04),PCHID=270
      CNTLUNIT CUNUMBR=0600,PATH=05,UNIT=3490,
               UNITADD=((00,16))
      IODEVICE ADDRESS=(600,16),CUNUMBR=0600,UNIT=3490,STADET=Y
*IOCP *
*IOCP * CHPID  FICON (vers baie de stockage DS8800)
*IOCP *
CHP06  CHPID   PATH=(CSS(0),06),TYPE=FC,SHARED,PCHID=118
CHP0A  CHPID   PATH=(CSS(0),0A),TYPE=FC,SHARED,PCHID=11C
CHP07  CHPID   PATH=(CSS(0),07),TYPE=FC,SHARED,PCHID=160
CHP0B  CHPID   PATH=(CSS(0),0B),TYPE=FC,SHARED,PCHID=161
      CNTLUNIT CUNUMBR=0200,PATH=(06,0A,07,0B),UNIT=3990,
               UNITADD=((00,128)),CUADD=0
      IODEVICE ADDRESS=(200,128),CUNUMBR=0200,UNIT=3390
      CNTLUNIT CUNUMBR=0300,PATH=(06,0A,07,0B),UNIT=3990,
               UNITADD=((00,32)),CUADD=2
      IODEVICE ADDRESS=(300,32),CUNUMBR=0300,UNIT=3390
*IOCP *
*IOCP *     Lecteurs cartouches 3590        *
*IOCP *
CHP0C  CHPID   PATH=(CSS(0),0C),TYPE=FC,SHARED,PCHID=119
CHP0D  CHPID   PATH=(CSS(0),0D),TYPE=FC,SHARED,PCHID=11D
      CNTLUNIT CUNUMBR=1600,PATH=((CSS(0),0C,0D)),
               UNITADD=((00,016)),UNIT=3590
      IODEVICE ADDRESS=(1600,016),CUNUMBR=(1600),STADET=Y,UNIT=3590
*IOCP *
*IOCP ***********************************************
*IOCP **     Fin des d{finitions                     **
*IOCP ***********************************************
```

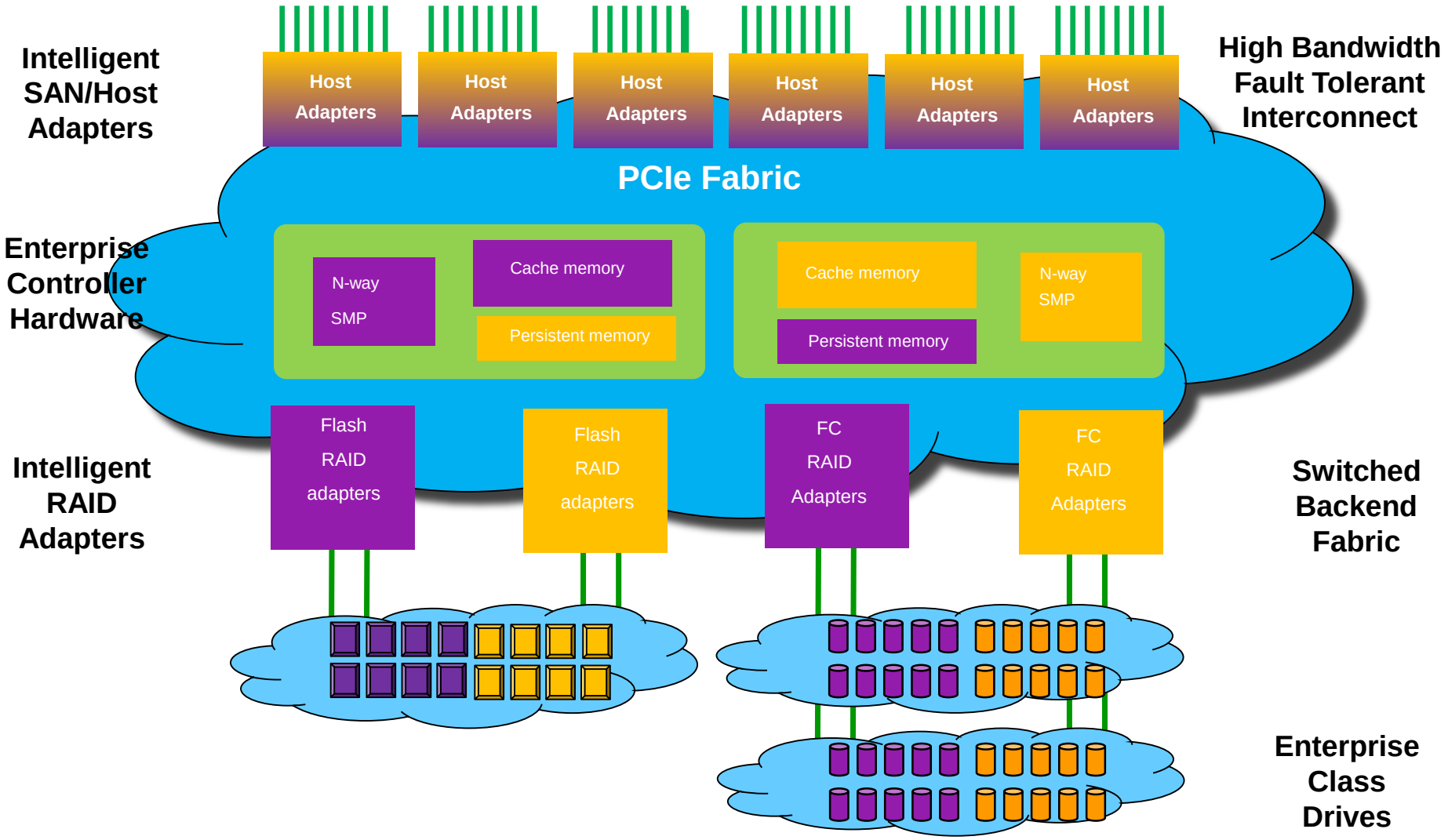
Le Service Element (SE)

Contrôle le fonctionnement de la CPU (arrêt, démarrage, diagnostique, IOCDS active …)

Permet de démarrer les partitions et les Operating System.

Les disques

DS8000 general HW structure



DS8880 family

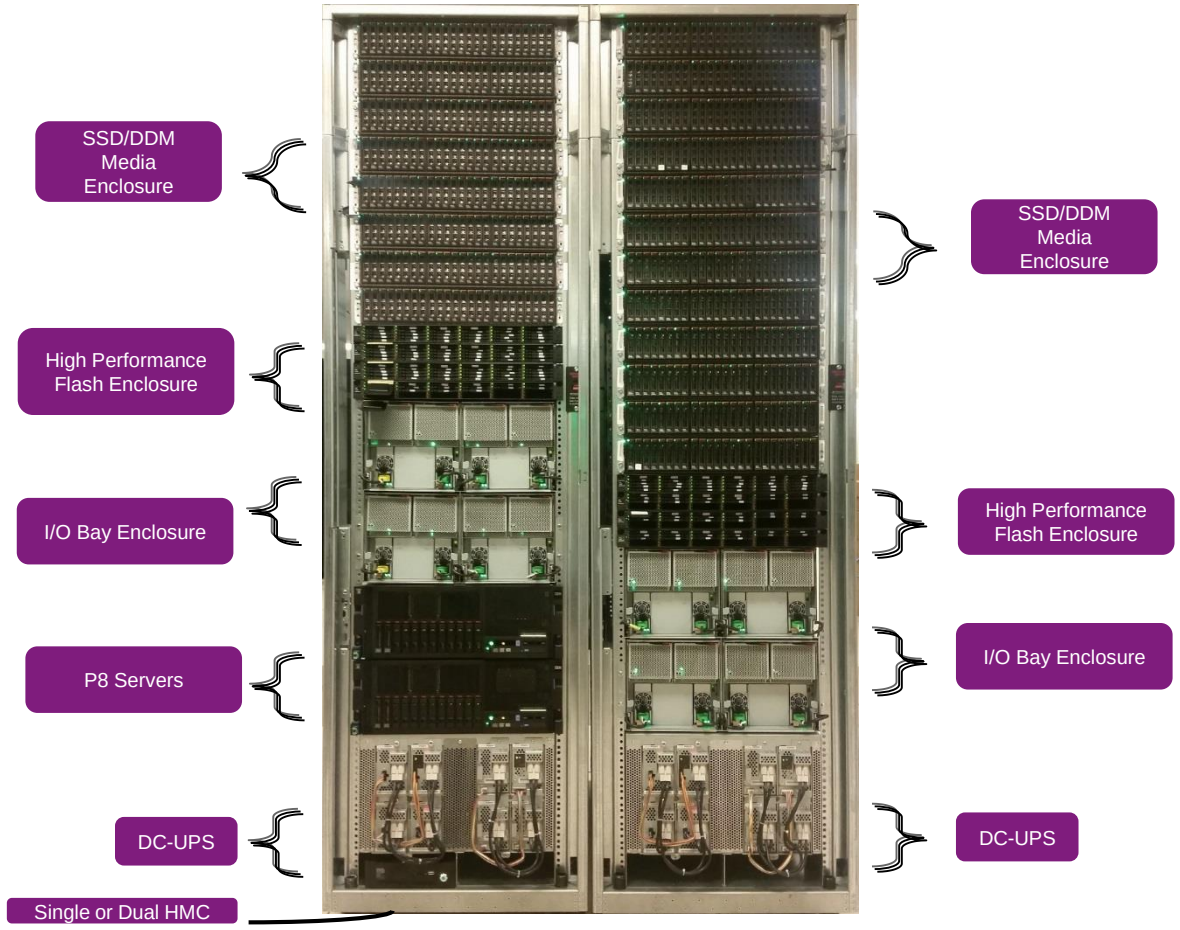
DS8 Advanced F(X)
Starting Price <$50K
256 GB Cache (DRAM)
64 Ports
768 HDD/SDD + 120 flash cards
19” Rack

DS8 Advanced F(X)
3X Performance
2 TB Cache (DRAM)
128 Ports
1536 HDD/SDD’s + 240 flash cards
19” Rack

DS8 Advanced F(X)
Industry’s Fastest T1 Subsystem
2 TB Cache (DRAM)
128 Ports
480 Flash Cards
19” Rack

DS8880

- Built on Power 8 Platform
- New IO Bay Interconnect
- PCIe Gen3 for increased Bandwidth performance
- 8 IOA PCIe connectors for increased HW support
- Single Phase Power
- Inherits all Advanced Function from DS8870
- Standard 19” rack
- Integrated Dual Hardware Management Controller (HMC)



Flash technology

	146GB 15k (RAID 10)	All-Flash (RAID 5)	All-Flash Benefit
Raw Capacity	152TB	90TB	41% Less
Usable Capacity	72TB	72TB	Same
Response Time	~ 1 ms	~ 0.3 ms	70% Less
Drive Count	1,056	224	80% Less
Frames	3 frames	2 frames	33% Less
Energy Usage	13.9kw	5.3kw	62% Less
$ / GB / IOP	Equivalent	Equivalent	Same

Drive technology

Performance

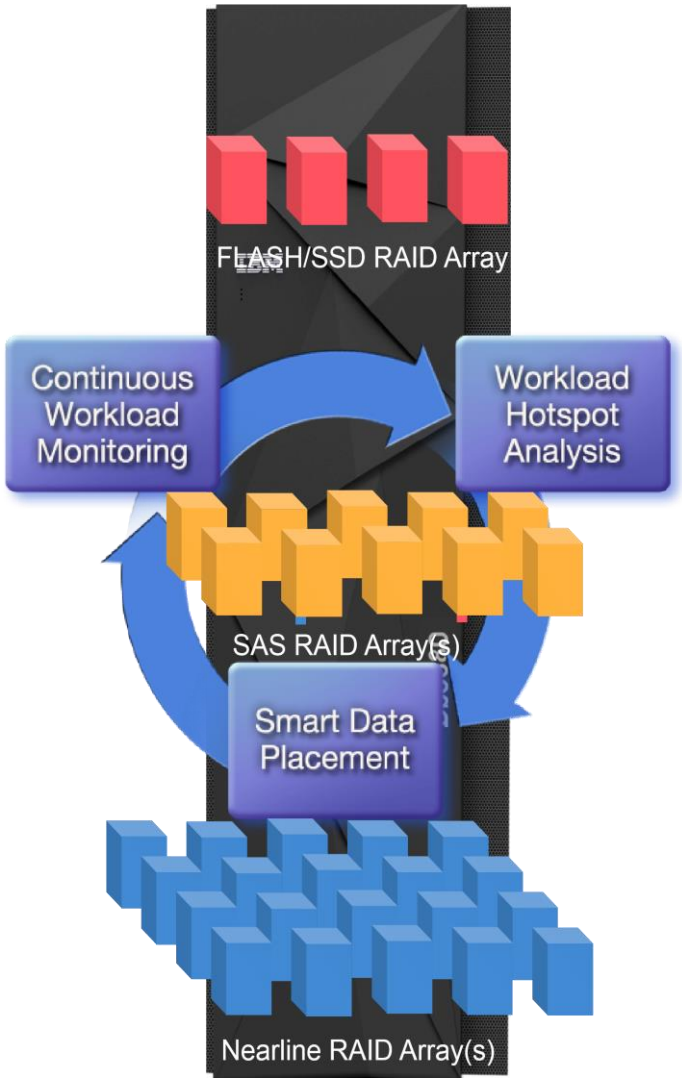
- Flash – 1.8” in High Performance Flash Enclosure
  - 400 GB drive, 800GB, 1.6TB, 3.2TB
- SSD – 2.5” Small Form Factor
  - Latest generation with higher sequential bandwidth
  - 200/400/800/1600GB SSD
- 2.5” Enterprise Class 15K RPM
  - Drive selection traditionally used for OLTP
  - 300/600GB drives
- 2.5” Enterprise Class 10K RPM
  - Large capacity, much faster than Nearline
  - 600/1200/1800GB drives
- 3.5” Nearline – 7200RPM Native SAS
  - Extremely high density, direct SAS interface
  - 4/6TB drives

DS8880 Easy Tier

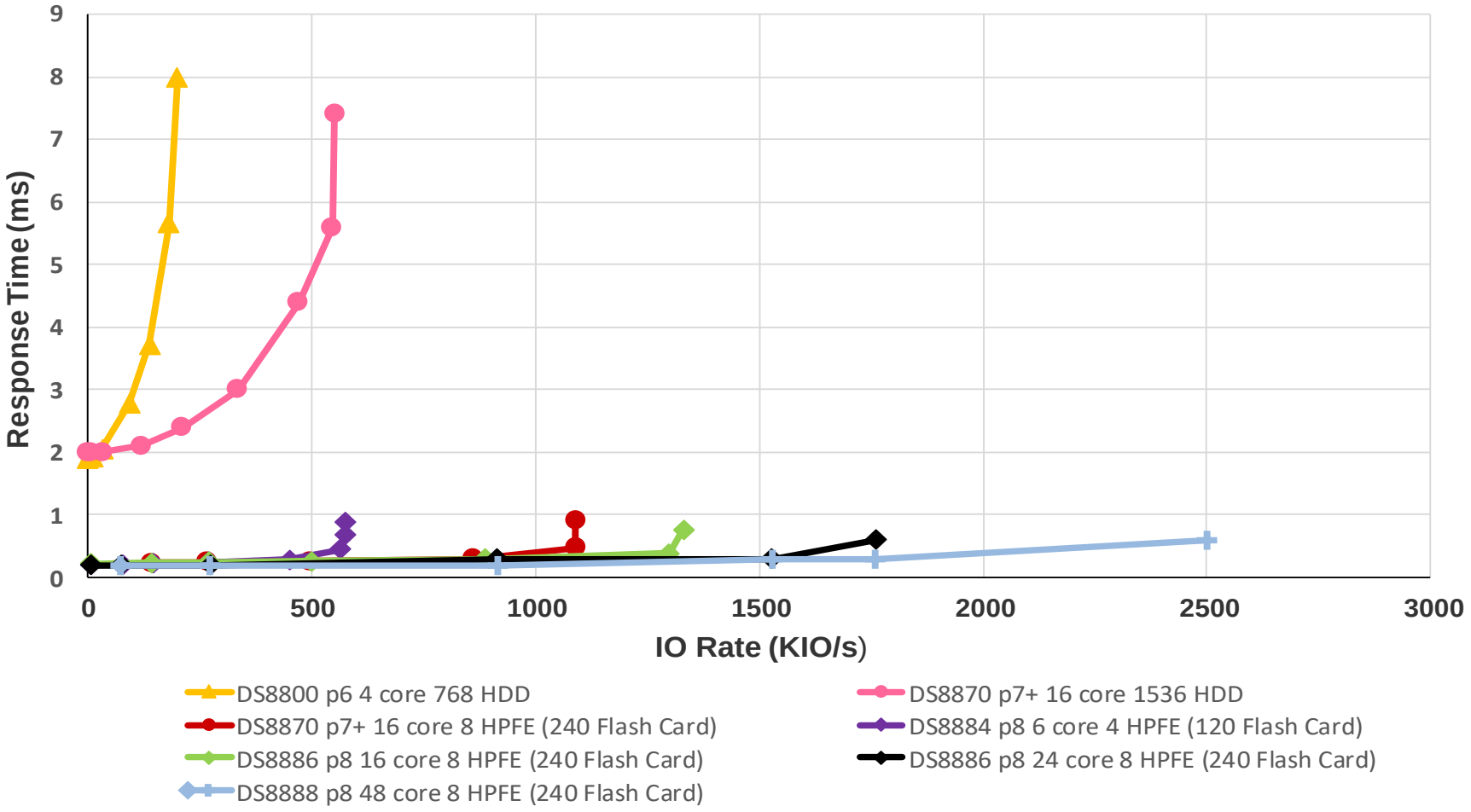
Easy Tier measures and manages activity
- Every five minutes: up to 8 extents moved
- New allocations placed initially on fastest HDD

Extent Pools can have mixed media
- Flash / Solid-State Drives (Flash / SSD)
- Enterprise HDD (15K and 10K RPM)
- Nearline HDD (7200 RPM)

As little as 3% Flash can dramatically reduce response times and increase IOPS



Up to 2.5M IOPS on z/OS OLTP



Storage Hierarchy

- Disk
  - Individual DDMs
- Array Sites
  - Logical Grouping of 8 DDMs of same speed and capacity and drive class
- Array
  - One 8-DDM Array Sites used to construct one RAID5, RAID-6 or RAID10 array
- Ranks
  - One Array becomes one CKD or FB Rank
  - Available space in rank divided into extents
    - An extent is the minimum allocation unit when a LUN or CKD volume is created (FB = 1GB, CKD = 1113 cylinders)
- Extent Pools
  - 1-N Ranks form an Extent Pool
    - Min of 2 pools—1 each for server0 and server1
    - Max of 1 pool for each rank
  - All Extents in a Pool are same storage type (CKD/FB); same RAID recommended
  - Associated with server0 or server1

EAV (Extended Address Volume)

zSeries IO operation are driven by channel program.

A channel program is a sequence of CCWs – Channel Command Word

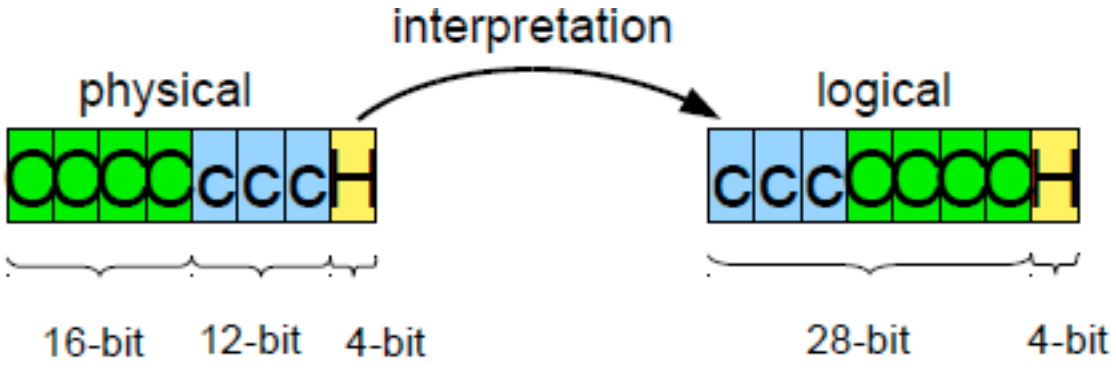
In ECKD architecture, to address a disk cylinder and head limits are:

- Cylinder 16 bits (from 0 to 65,535)
- Head 16 bits (from 0 to 65,535)

In EAV architecture, limits are:

- Cylinder 24 bits (from 0 to 268,435,455)
- Head 4 bits (from 0 to 15)

Currently up to 262668 cylinders (approximately 223.2 GB raw, 180GB formatted)



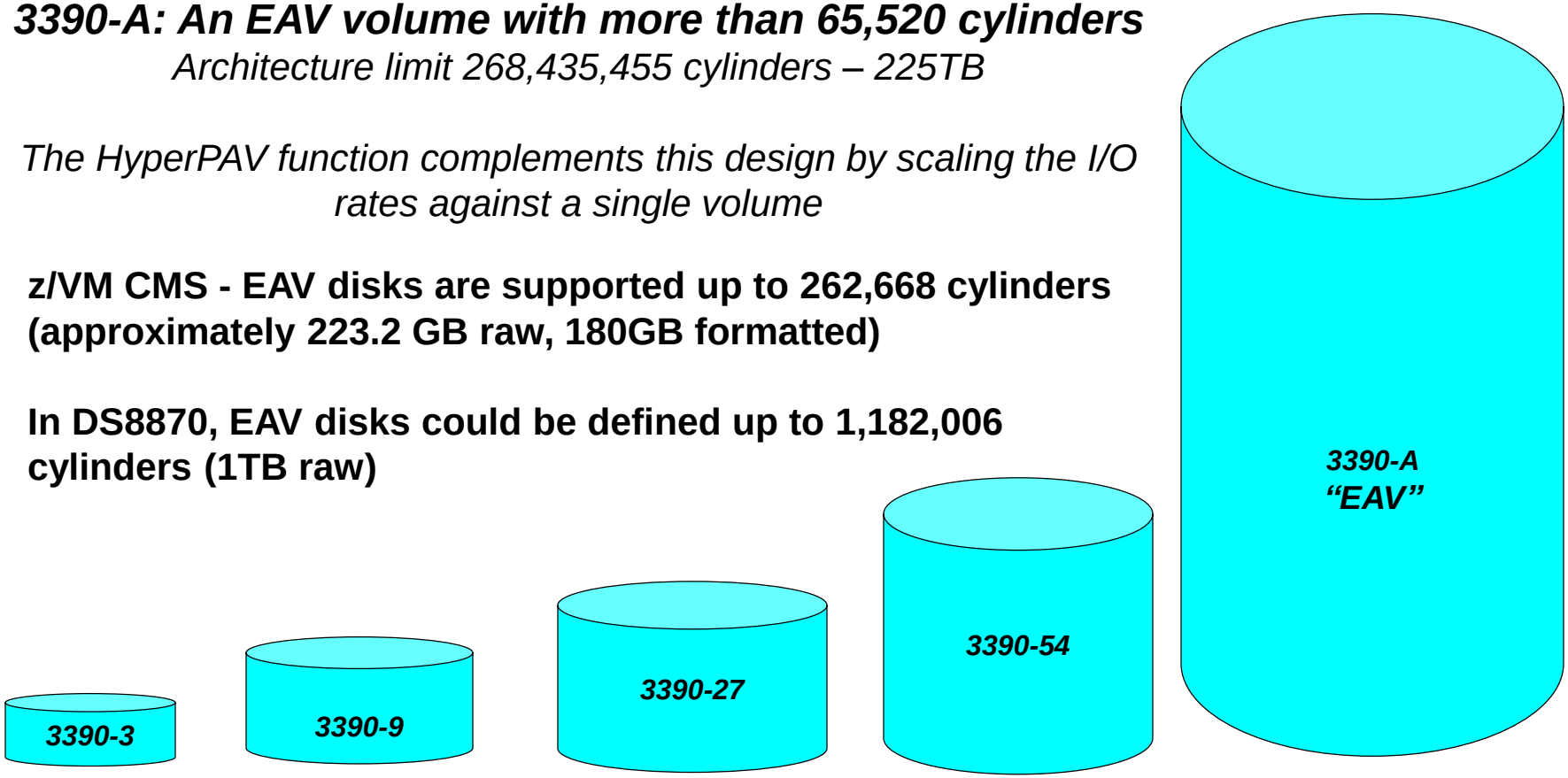
3390-A: An EAV volume with more than 65,520 cylinders

Architecture limit 268,435,455 cylinders – 225TB

The HyperPAV function complements this design by scaling the I/O rates against a single volume

z/VM CMS - EAV disks are supported up to 262,668 cylinders (approximately 223.2 GB raw, 180GB formatted)

In DS8870, EAV disks could be defined up to 1,182,006 cylinders (1TB raw)



3GB	9GB	27GB	54GB	1 TB
Max cyls: 3,339	Max cyls: 10,017	Max cyls: 30,051	Max cyls: 60,102	Max cyls: 1,182,006

Maximum Sizes

Disaster recovery

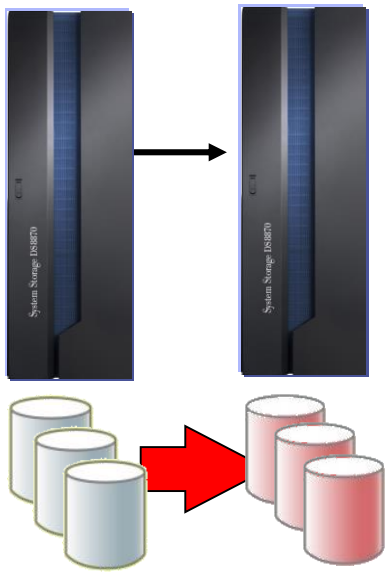
FlashCopy
Point in time copy

Within the same Storage System

Metro Mirror
Synchronous mirroring

Primary Site A

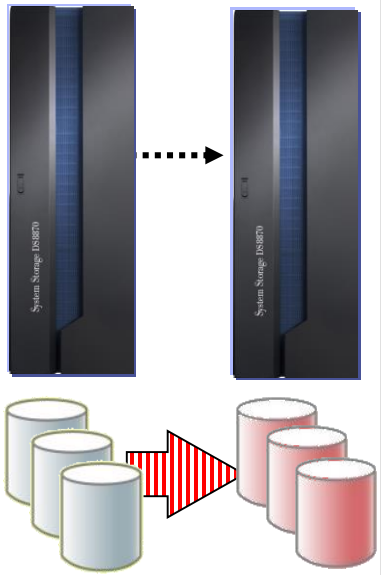
Metro distance Site B



Global Mirror
z/OS Global Mirror
Asynchronous mirroring

Primary Site A

Out of Region Site B

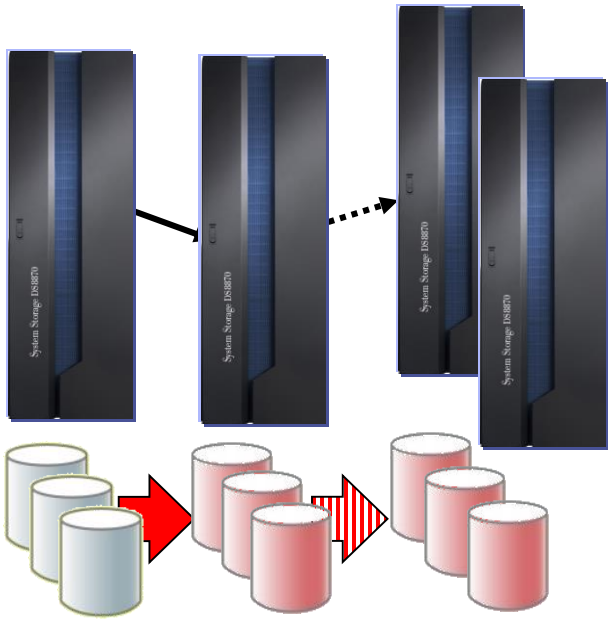


Metro Global Mirror
Metro z/OS Global Mirror
Three site and Four Site synchronous & asynchronous mirroring

Primary Site A

Metro Site B

Out of Region Site C/D

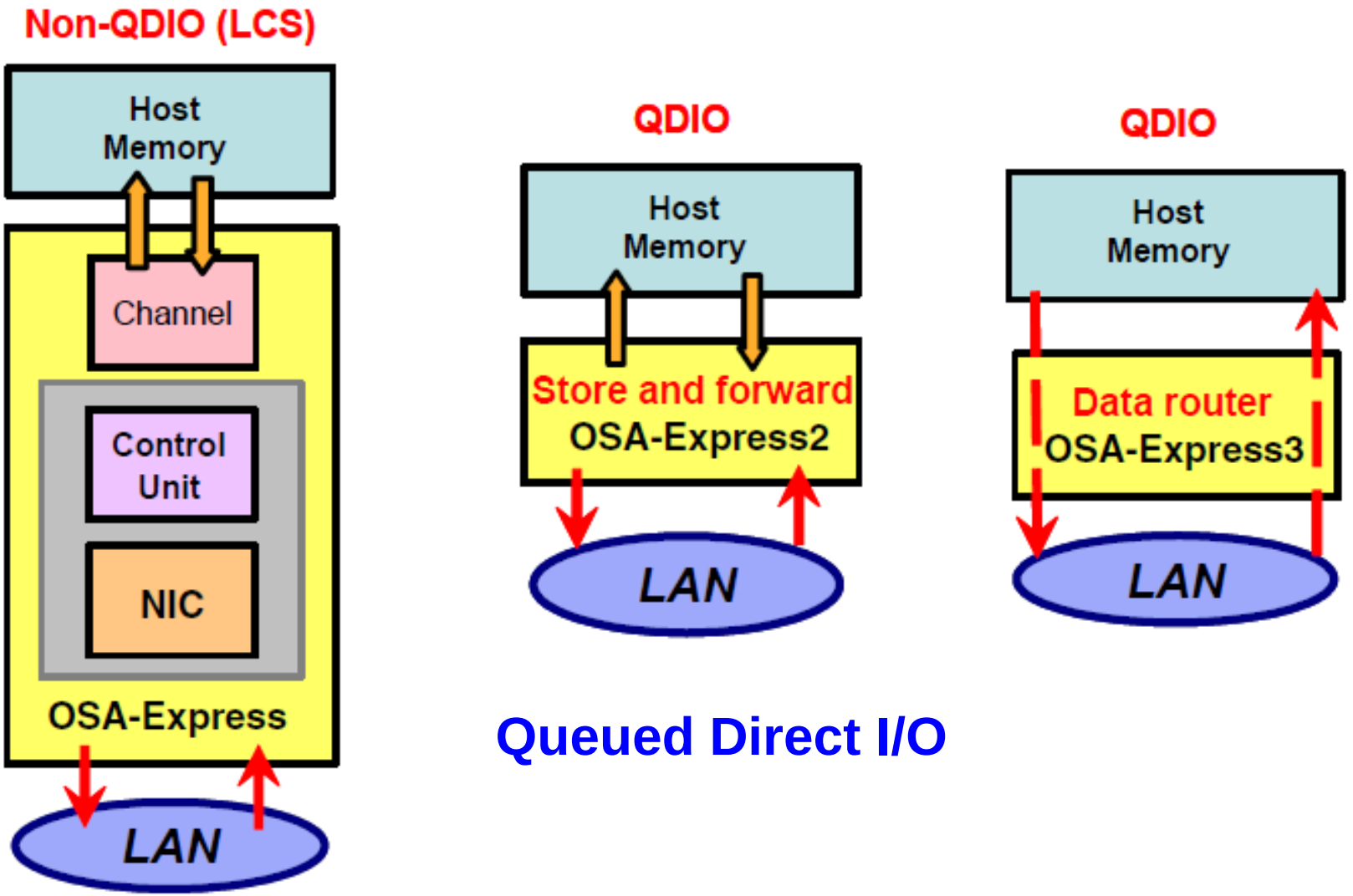


Les cartes OSA

Les cartes OSA

- OSA Express 5S model
  - 1000baseT rj45 2 interfaces 1Gbit full duplex
  - GbE fiber 2 interfaces 1Gbit full duplex
  - 10GbE fiber 1interface 10Gbit full duplex
- Shared between up to 640 TCP/IP stacks
- Three devices (I/O subchannels) per stack:
  - Read device (control data <-- OSA)
  - Write device (control data --> OSA)
  - Data device (network traffic)
- Network traffic Linux <--> OSA
  at IP (layer3) or Ethernet (layer2) level
- One MAC address for all stacks (layer 3)
- OSA handles ARP (layer 3)
  (Address Resolution Protocol)

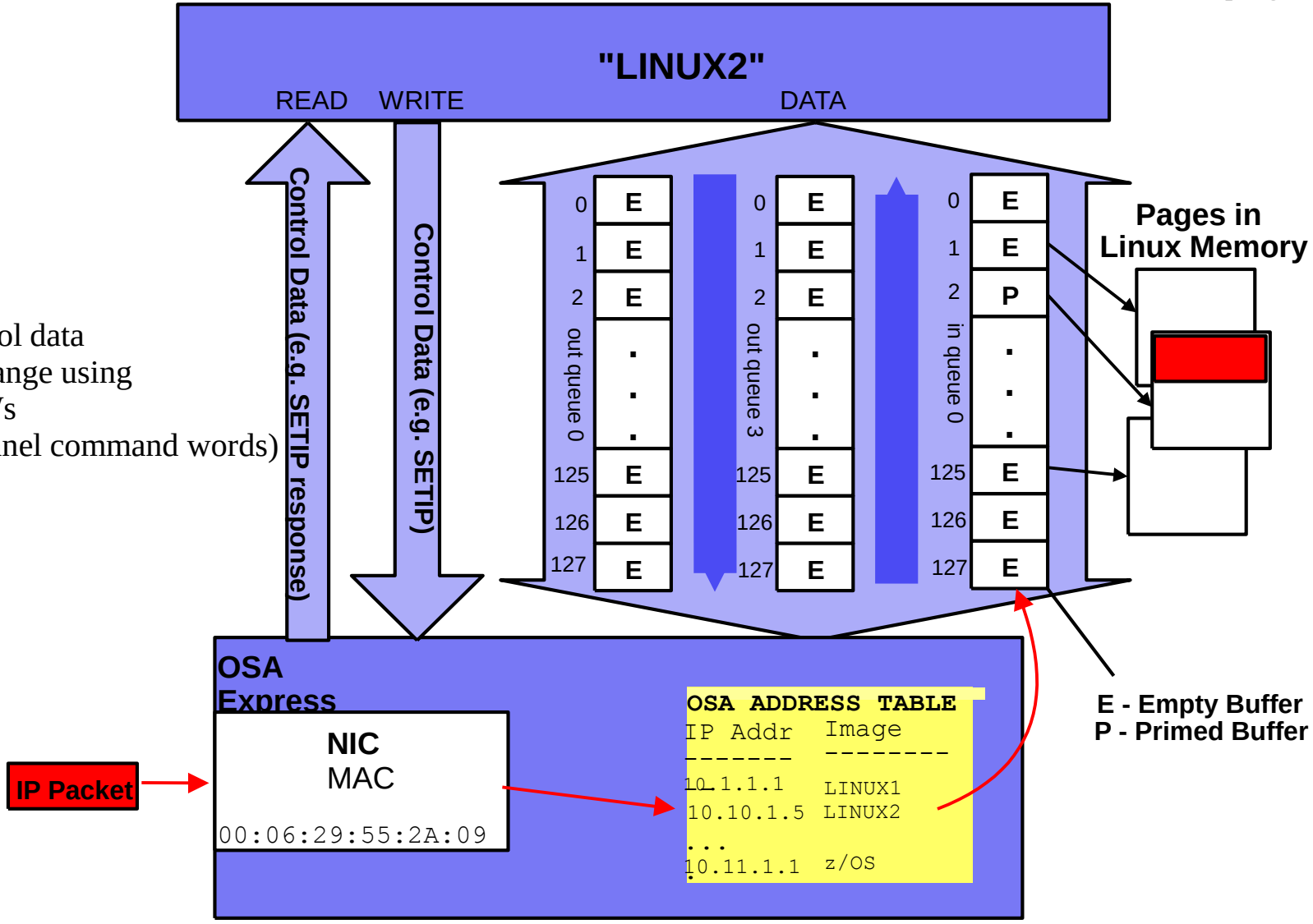
OSA card Mode LCS ou QDIO



Queued Direct I/O

The Queued Direct I/O ⇔QDIO

QDIO as “long-running” channel program



Les dispositifs des cartes OSA

- VIPA
- Large Send
- VLAN
- SNMP support
- ARP
- Layer 2 support
- Link Aggregation in layer 2
- QDIO data isolation (VSWITCH port isolation)
- OSA/SF (z/VM)

Les bandes

TS4500 Tape Library Overview

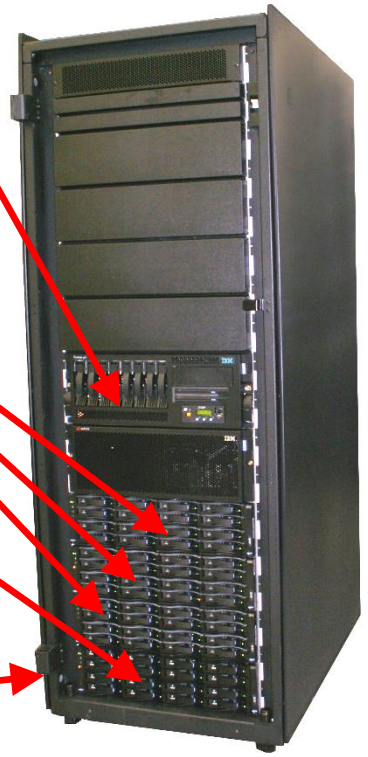
- Up to 17 expansion frames
  - Frames for LTO drives and media
  - Frames for TS1140 / TS1150 drives and media
- Up to 128 tape drives
- Storage Capacity of up to 175.5 PB per library (526.5 PB with 3:1 compression)

TS1140 Tape Drive

- 4th Generation of 3592 enterprise tape drive roadmap
  - 250 MB/sec performance (up to 800 MB/s at 3:1 compression)
  - 4TB (using JC/JY media), 1.6TB (using JB/JX media or 500GB (using JK media)
  - Supports WORM cartridges and data encryption
- Supported in
  - IBM TS4500 tape libraries

TS7740 Virtualization Engine Components

- TS7740 Virtualization Engine (3957 Model V07)
  - Power7 Atlas HV32 Server
  - One 3Ghz 8 processor core card, 16GB Memory
  - Runs the V and H nodes (Gnode)
  - Integrated Enterprise Library Controller
- TS7740 Cache Drawer (3956 Model CX7)
  - 16 15K 600GB FC HDDs
  - 7.04 TB usable capacity (after RAID 5 and spares)
  - Three maximum drawers – Total 21.12 TB additional
- TS7740 Cache Controller (3956 Model CC8)
  - Disk RAID array controller
  - 16 15K 600 GB FC HDDs
  - 7.04 TB usable capacity (after RAID 5 and spares)
  - Adding up to three CX7 drawers provides up to 28.17 TB total capacity
- 3952 Model F05 Frame
  - Houses major components & support components
  - Dual Power



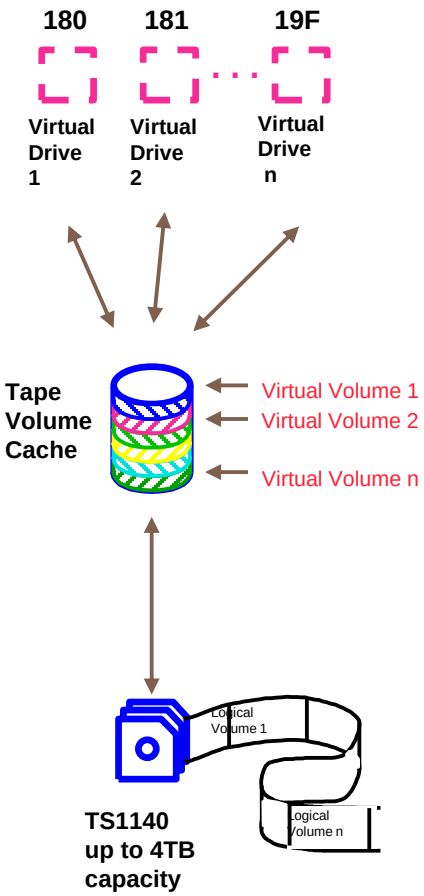
The combination of the Virtualization Engine components is called a TS7700 Cluster

Virtual Tape Server

- **Virtual tape drives**
  - Appear as multiple 3490E tape drives
  - Shared / partitioned like real tape drives
  - Designed to provide enhanced job parallelism
  - Requires fewer real tape drives
  - TS7700 offers 256 virtual drives per cluster

- **Tape volume caching**
  - All data access is to disk cache
  - Removes common tape physical delays
  - Fast mount, positioning, load, demount
  - Up to 28 TB / 440 TB of cache (uncompressed)

- **Volume stacking (TS7740)**
  - Designed to fully utilize cartridge capacity
  - Helps reduces cartridge requirement
  - Helps reduces footprint requirement

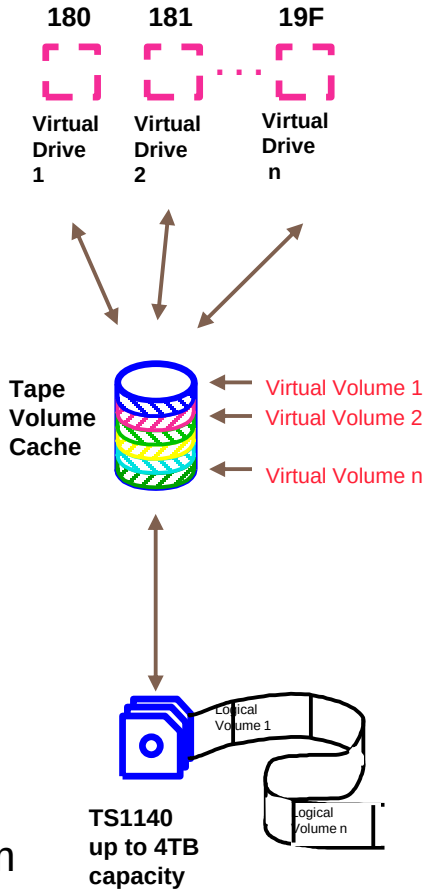


Virtual Tape Server

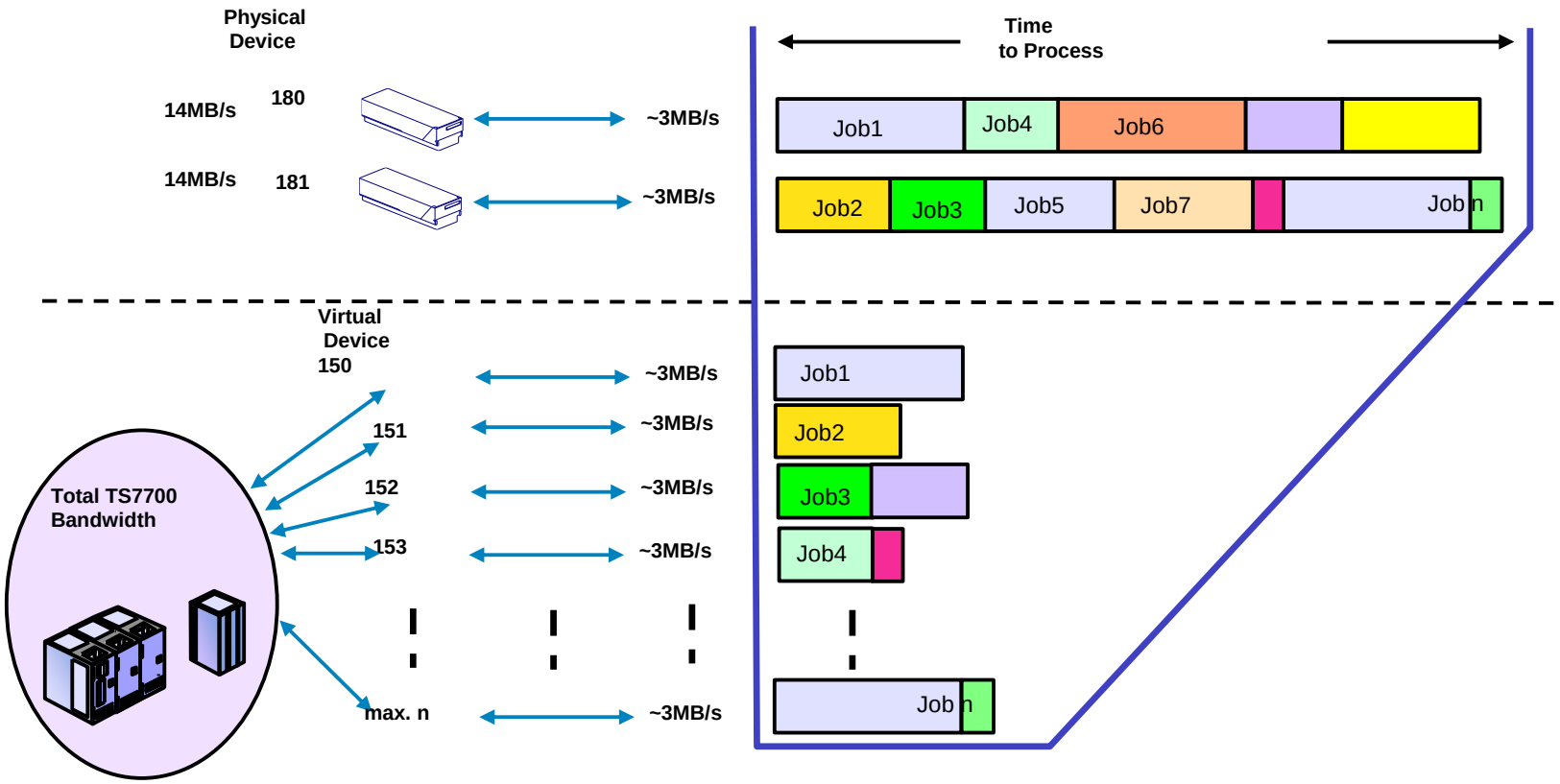
- **Virtual volume copy function**
  - Maximize cache hits
  - Keep cache full with virtual volumes
  - Provide free space if required
    - Implemented through (pre-)copy process
    - Logical Volumes are copied in FIFO order
    - Scheduled after End of Volume (EOV) completes
    - Copy process asynchronous from host process

- **Optional cache space management**
  - SMS preference groups PG0/PG1
  - Uses Last Recently Used (LRU) algorithm

- **Automatically manages physical cartridge space**
  - Reclaims ‘gaps’ caused by logical volume expiration
  - User set policies to control when a cartridge is eligible for reclaim

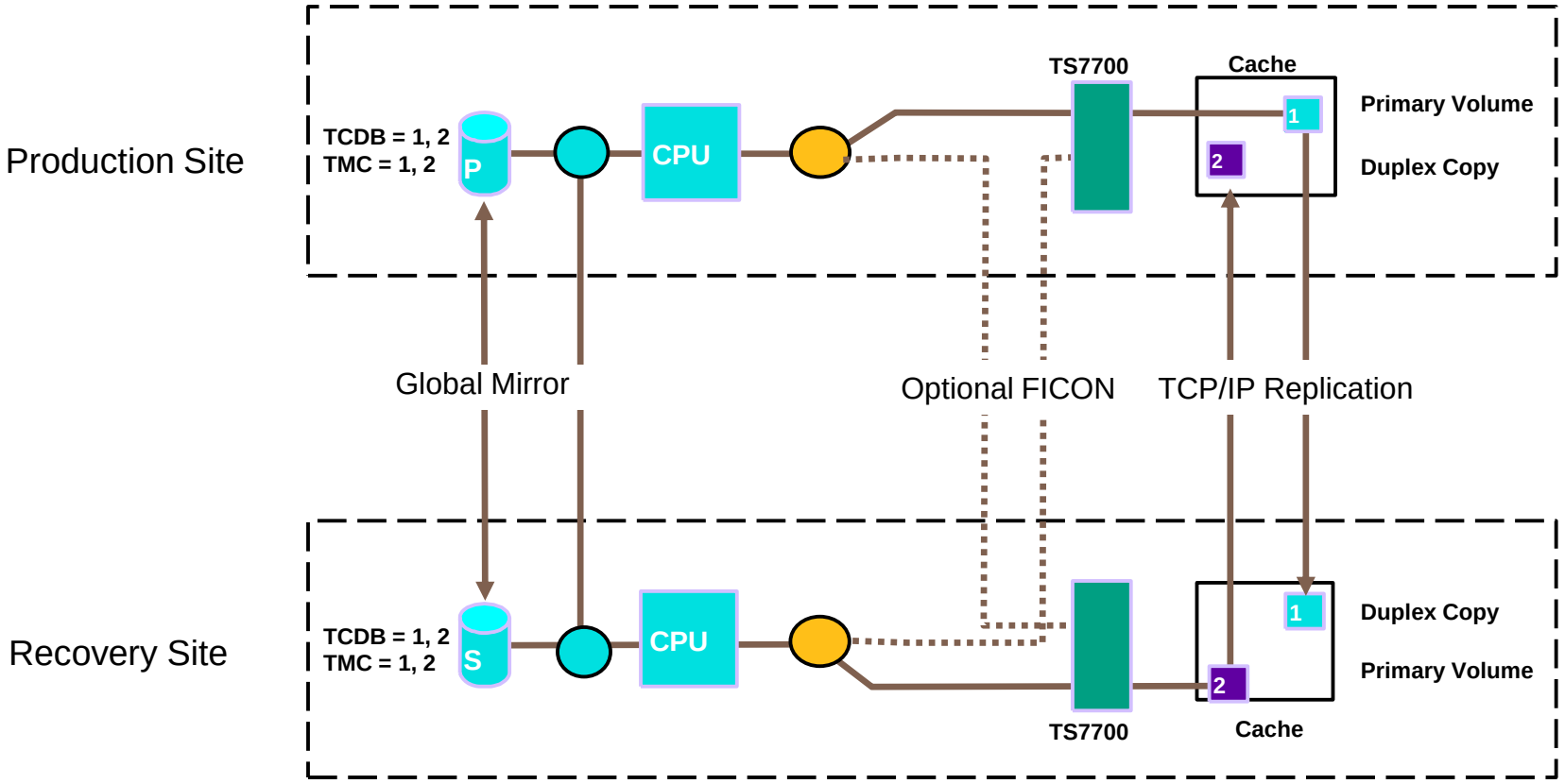


TS7700 Virtualization Engine Performance



Peak Performance is a key measurement criteria

Supports Tier 5 Business Continuity



This diagram is meant to illustrate the logical process of the primary and secondary virtual volume copy creation, and does not include all steps in the process.