Maciej Przepiórka
Systems Architect, IBM Polska

POWER7
Zaawansowane technologie serwerowe

Wartości systemów IBM Power

Wirtualizacja state-of-the-art

- utylizacja serwerów do 90%
- PowerVM w 65% systemów w 2009

Efektywność energetyczna

- redukcja kosztów o 70-90%
- wysoka wydajność / watt

AIX® i for Business Linux®

Wydajność, skalowalność, stabilność, bezpieczeństwo

Wysoka ciągłość pracy

- zaawansowane technologie RAS
- oprogramowanie PowerHA (/ XD)

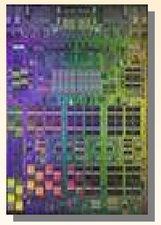
IBM® Systems Director

Łatwość zarządzania

- większa szybkość wdrożeń
- kontrola zużycia energii i kosztów

Rozwój procesorów POWER

2001
POWER4™



180 nm

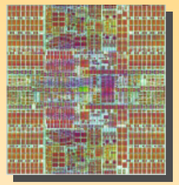
- **Wysoka wydajność rdzenia**
- **Pierwszy procesor dwurdzeniowy na świecie**
- **Architektura wieloprocesorowa**
- **Distributed Switch**
- **Współdzielona pamięć L2**
- **Skalowalność do 32 rdzeni**
- **Dynamiczne partycje LPAR**

2004
POWER5™

130 nm

- **Wyższa wydajność rdzenia**
- **Skalowalność do 64 rdzeni**
- **Simultaneous Multi-Threading**
- **Enhanced Distributed Switch**
- **Zintegrowany kontroler pamięci**
- **Niższe czasy dostępu do pamięci L3 i RAM**
- **Wirtualizacja i mikropartycje**

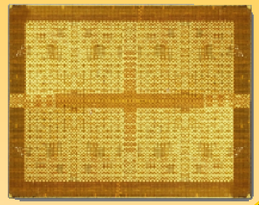
2007
POWER6™



65 nm

- **Wyższa wydajność rdzenia**
- **Bardzo wysoka częstotliwość taktowania (do 5 GHz)**
- **Usprawniony tryb SMT**
- **Zarzadzanie energią**
- **Rozrzerzenia wirtualizacji**
  - **Processor Pooling**
  - **Partition Mobility**
- **Usprawniony kontroler pamięci**
- **Instrukcje Altivec / SIMD**
- **Zaawansowane funkcje RAS**

2010
POWER7™



45 nm

- **Wyższa wydajność rdzenia**
- **Więcej rdzeni/socket**
- **Skalowalność do 256 rdzeni**
- **Usprawniony tryb SMT**
- **Zaawansowany kontroler pamięci DDR3**
- **Pamięć podręczna L3 eDRAM**
- **Rozszerzone funkcje obsługi pamięci**

pełna zgodność binarna

Nowości w POWER7

Skalowalność i wydajność

- procesor POWER7 jest układem 8-rdzeniowym
- każdy rdzeń może pracować w 4 wątkowym trybie SMT
- łącznie daje to 32 wątki na procesor (chip), czyli 8 razy więcej niż w POWER6
- *Intelligent Threads* – liczba wątków SMT jest dynamicznie zmieniana w zależności od charakterystyki aplikacji
- technologia *TurboCore* – zwiększamy częstotliwość zegara kosztem liczby rdzeni, uzyskując większą wydajność aplikacji jednowątkowych (L3 pozostaje nadal 32MB)

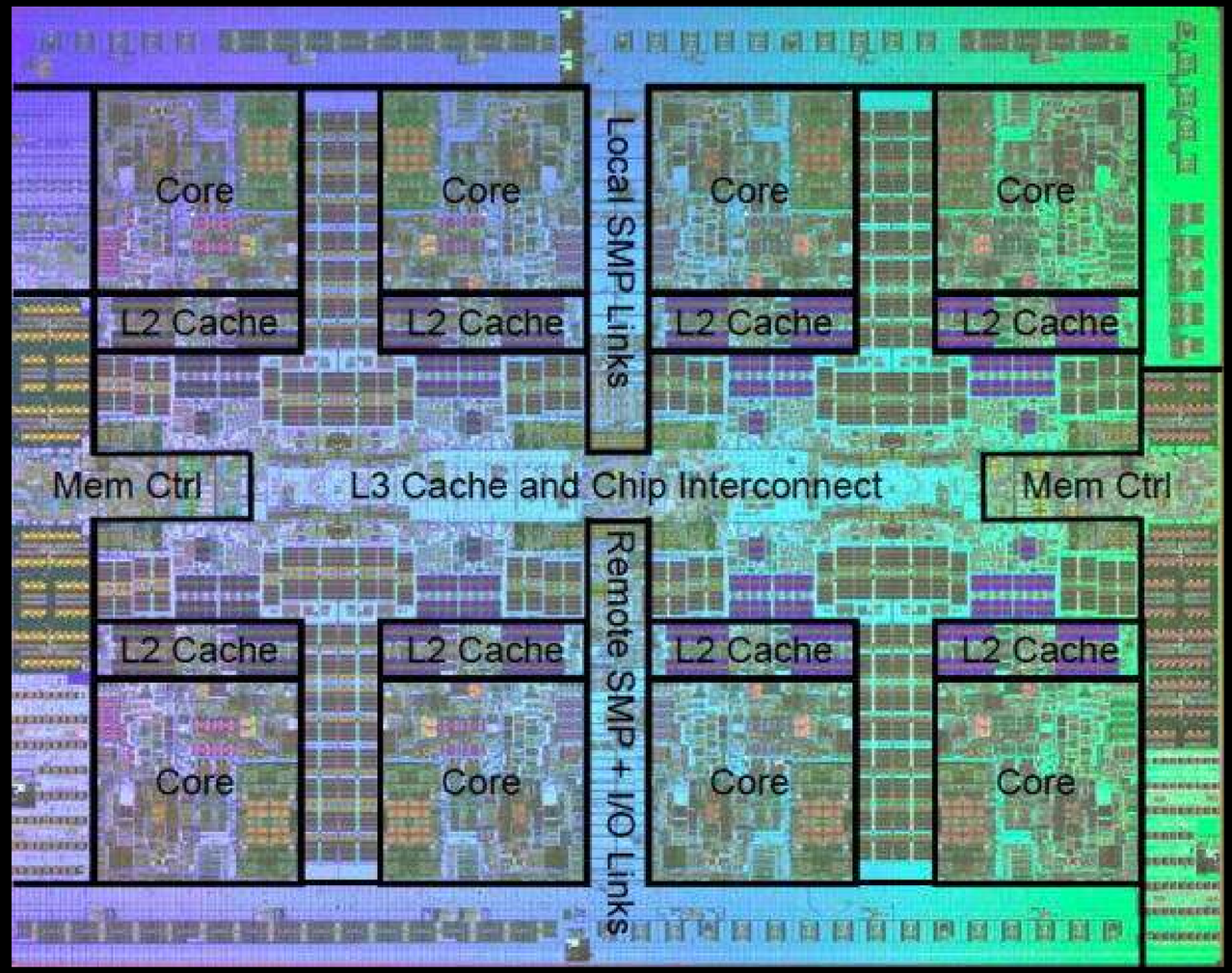
Wirtualizacja

- wsparcie dla 1000 partycji logicznych (maszyn wirtualnych) na jeden serwer fizyczny
- Active Memory Expansion – kompresja pamięci RAM w locie, w sposób przezroczysty dla aplikacji

Energooszczędność

- dynamiczne wyłączanie nieużywanych rdzeni lub obniżanie częstotliwości zegara w zależności od obciążenia

Procesor POWER7 w powiększeniu

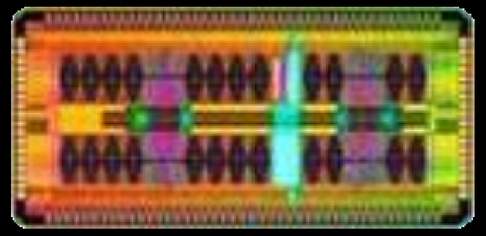


Zgodność binarna z poprzednimi generacjami procesorów POWER.

- Liczba rdzeni: 8 (opcjonalnie 4 lub 6)
- Powierzchnia 567 mm^2
- Technologia litograficzna 45 nm
- Liczba tranzystorów: 1,2 miliarda
  - tak „niewielka” dzięki zastosowaniu pamięci eDRAM
  - równowartość 2.7 miliarda w przypadku zastosowania konwencjonalnej pamięci cache
- Osiem rdzeni procesorowych
  - 12 jednostek wykonawczych na rdzeń
  - 4-wątkowy tryb SMT
  - 32 wątki na układ procesorowy
  - L1: 32 KB I Cache / 32 KB D Cache
  - L2: 256 KB na rdzeń
  - L3: współdzielone 32MB eDRAM na procesorze
- Dwa kontrolery pamięc DDR3
  - przepustowość do pamięci RAM: 136 GB/s
- Skalowalność do 32 układów zapewniają połączenia I/O SMP o przepustowości 360 GB/s na każdy procesor

Technologia eDRAM

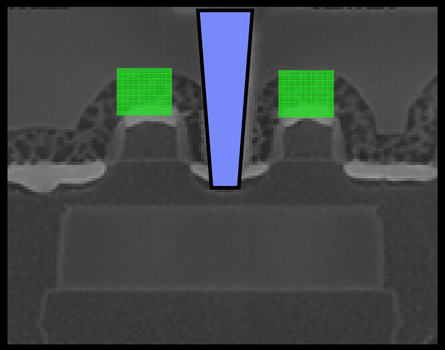
Procesor POWER7 posiada 32 MB pamięci cache 3-ciego poziomu zbudowanej na własej strukturze krzemowej.



Układ eDRAM

Zalety pamięci eDRAM (embedded Dynamic RAM) nad pamięcią SRAM (Static RAM):

- 3 razy większa gęstość komórek (bitów)
- 5-krotnie niższy pobór energii
- 250-krotnie rzadsze występowanie błędów



Bit eDRAM

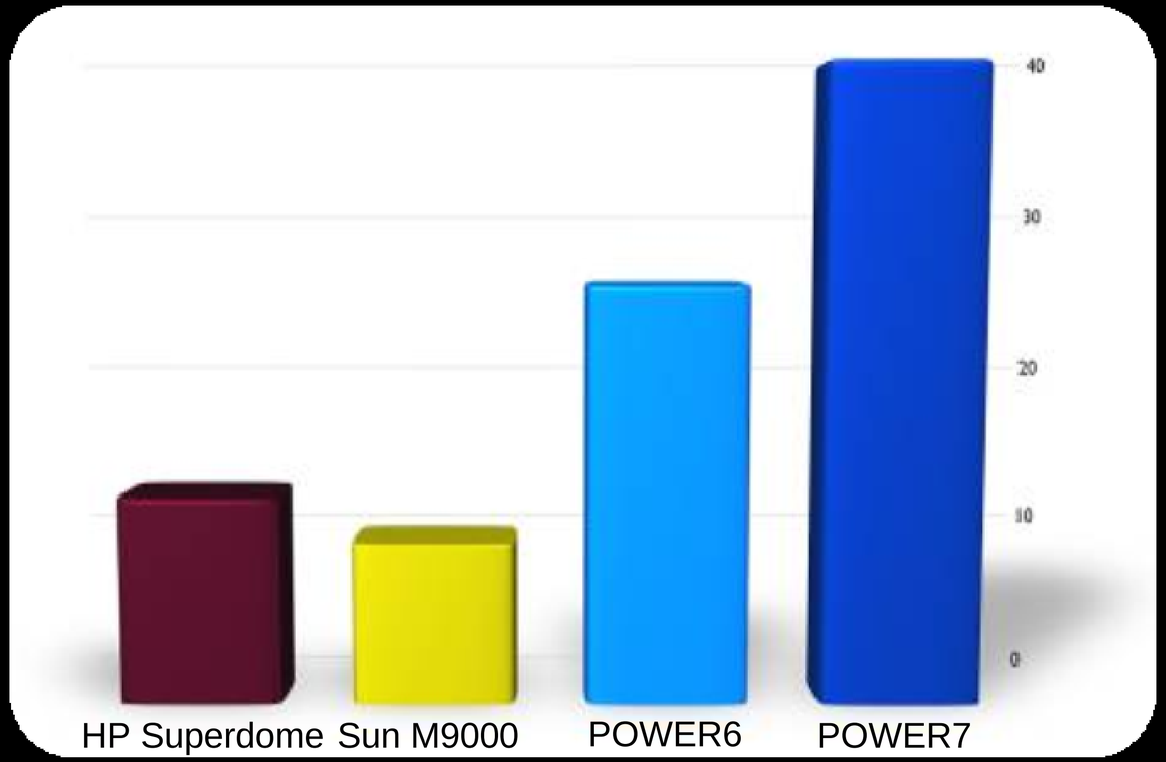
Zalety wbudowanej pamięci cache L3 nad zewnętrzną (jak w POWER6):

- 6-krotnie krótsze oczekiwanie na dane
- 2-krotnie wyższa przepustowość

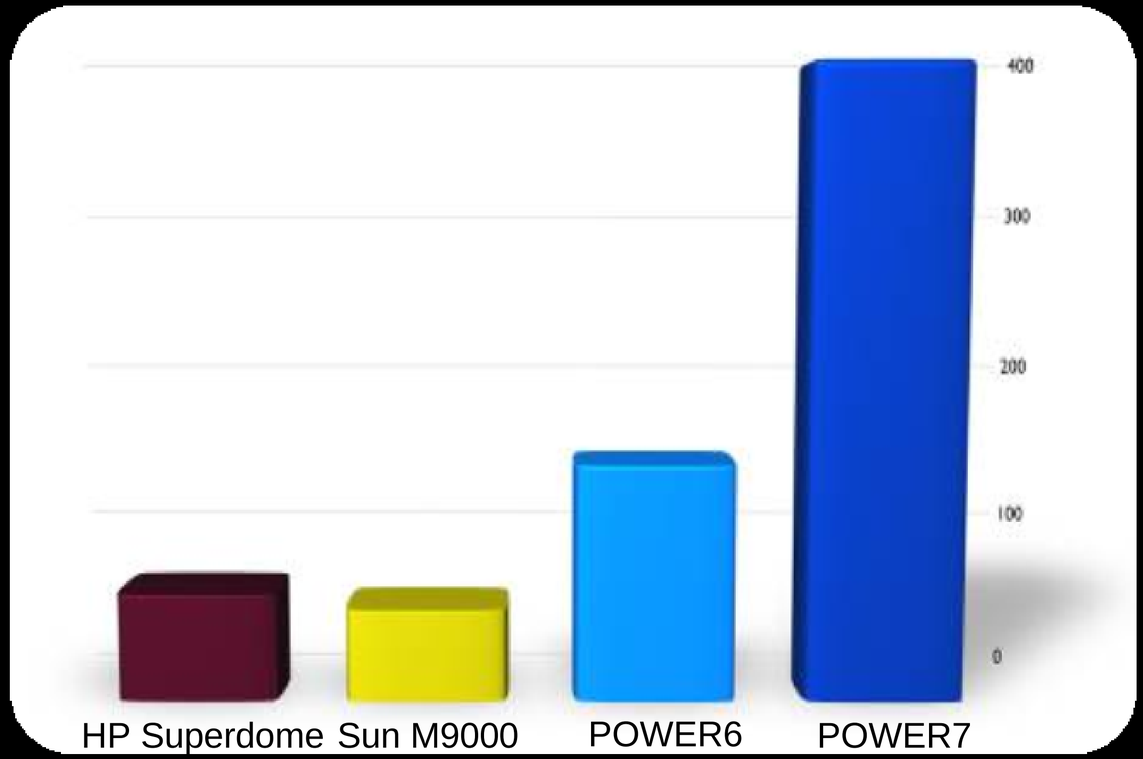
Dzięki eDRAM, procesor POWER7 osiąga wyższą wydajność przy jedniczesnej redukcji liczby tranzystorów o 1,5 miliarda oraz osiąga niespotykaną wydajność w przeliczeniu na jednostkę zasilania.

POWER7 – wydajne rdzenie o niskim poborze mocy

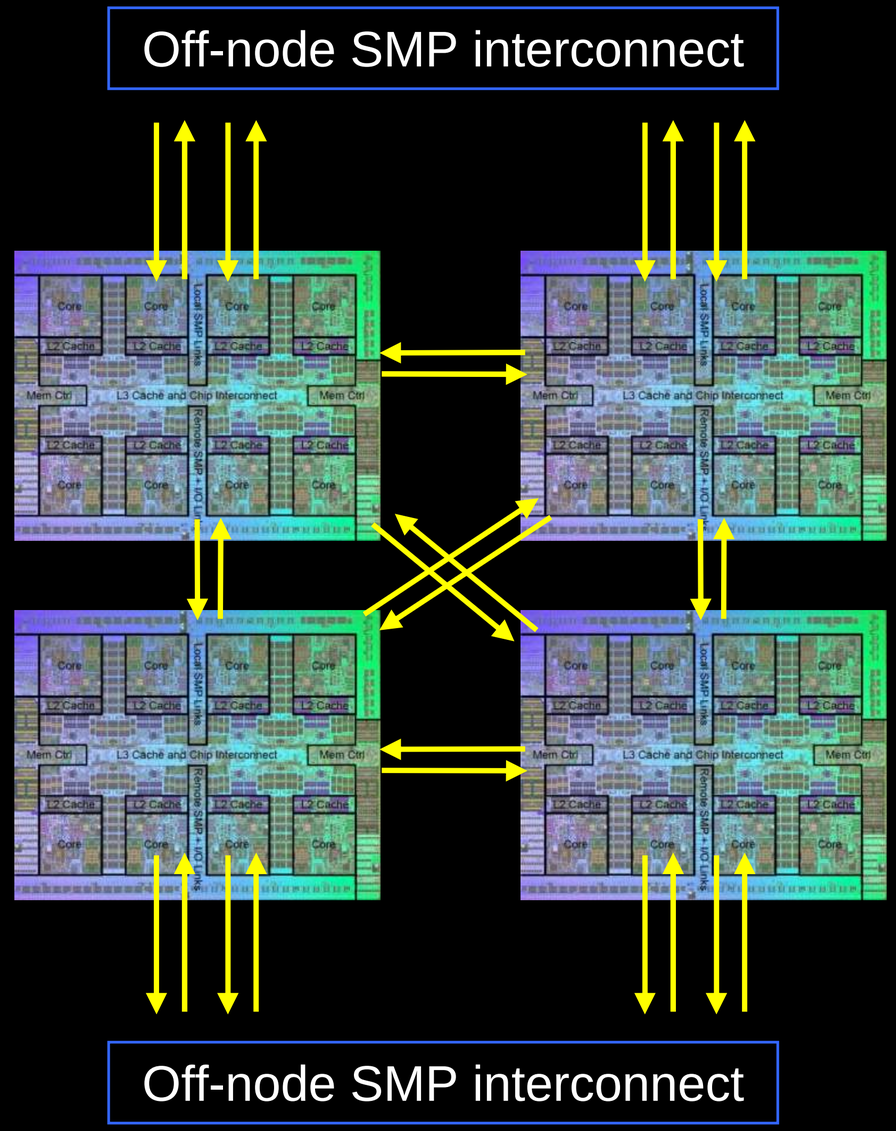
Performance per core (SPECint)



Performance per Watt



Jak osiągamy dobrą skalowalność?



Moduł wieloprocesorowy oparty o POWER7:

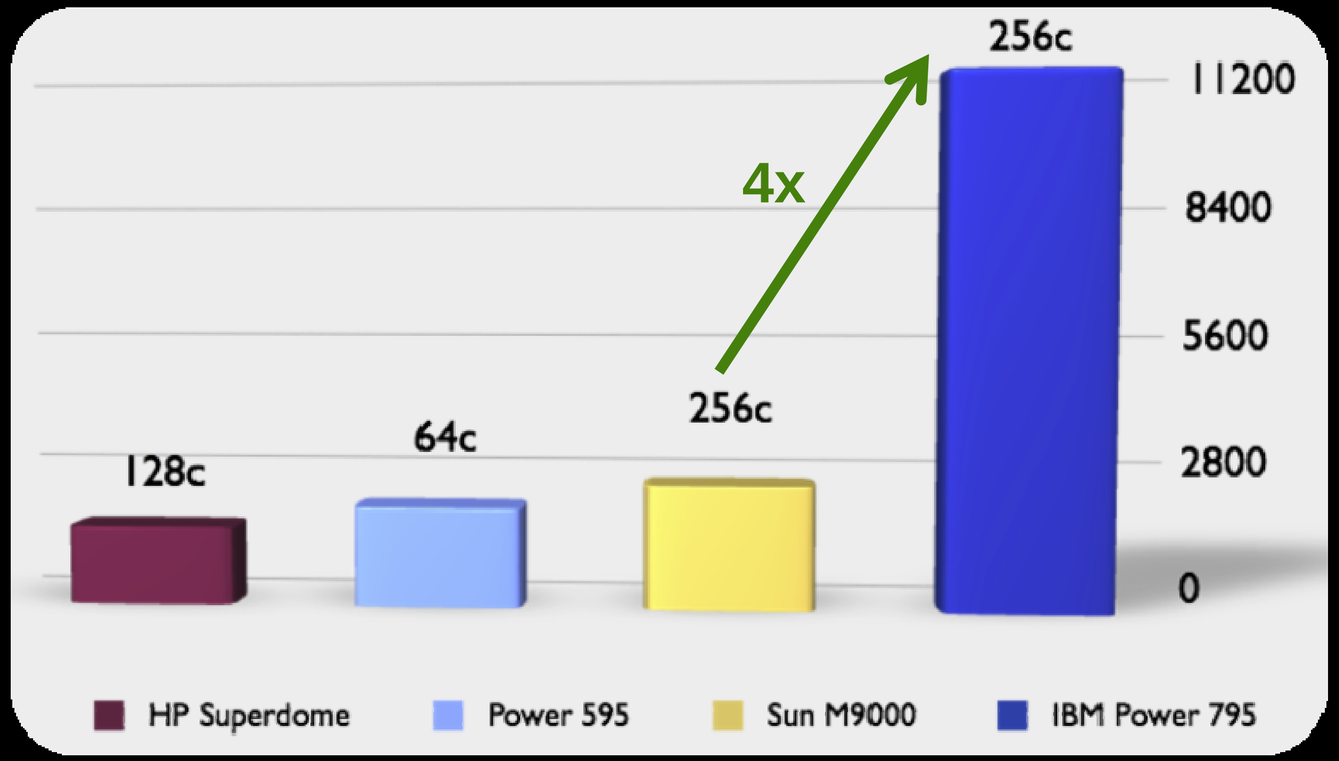
- łącznie 4 procesory (32 rdzenie)
- niezależne nieblokujące się połączenia
  - przepustowość 360 GB/s na każdy procesor (8 rdzeni)
- łączna przepustowość do pamięci RAM:
  - do 540 GB/s na każde 32 rdzenie
  - pamięć lokalna i zdalna – NUMA, ale zachowując spójność danych (SMP)
- zwiększając liczbę procesorów jednocześnie zwiększamy przepustowość do pamięci RAM

Power 795 – skalowalność klasy enterprise

Memory	Bandwidth
L1 (Data)	= **192** GB/sec
L2	= **192** GB/sec
L3	= **128** GB/sec
Memory 8 Nodes	**136.448 GB/sec per socket** **Total = 4366.366 GB/sec**
Intra-Node Buses	= **120** GB/sec per Socket = **480** GB/sec per Node = **3840** GB/sec per System
Inter-Node Buses	= **26.7**GB/sec 28 * 26.7 = **746.7** GB/sec per System
GX Bus	= **20** GB/sec 640 GB/sec per System

Power 795 – wydajność

SPECint_rate

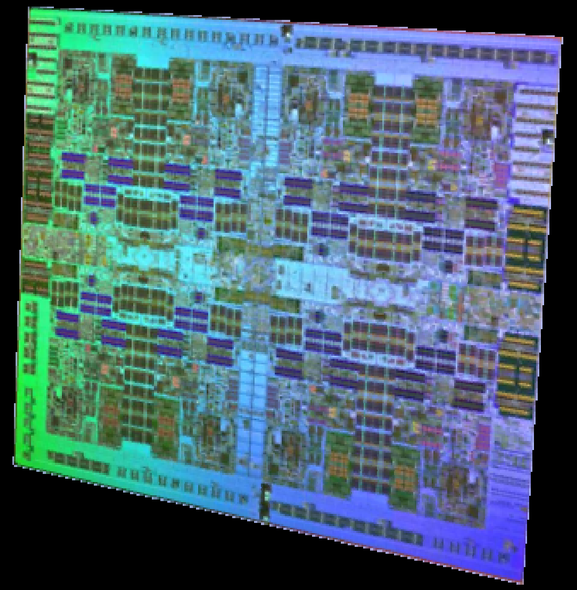


Wydajność baz danych na systemach IBM Power

- Numer 1 w TPC-C w systemach nieklastrowanych:
  Power 595, 64 rdzenie POWER6 5.0 GHz: **6M tpmC**
- Numer 1 w TPC-C w systemach klastrowanych (DB2 pureScale):
  Power 780, 192 rdzenie POWER7 3.8 GHz: **10.3M tpmC**
- Numer 1 w TPC-C w wydajności per core:
  Power 780, 8 rdzeni POWER7 4.14 GHz: **1.2M tpmC (150k tpmC per core)**

Rank	Company	System	Performance (tpmC)	Price/tpmC	Watts/KtpmC	System Availability	Database	Operating System	TP Monitor	Date Submitted	Cluster
1	IBM	IBM Power 780 Server Model 9179-MHB	10,366,254	1.38 USD	NR	10/13/10	DB2 9.7	AIX Version 6.1	Microsoft COM+	08/17/10	Y
2	ORACLE	Sun SPARC Enterprise T5440 Server Cluster	7,646,486	2.36 USD	NR	03/19/10	Oracle Database 11g Ent. Ed. w/Real Application Clusters w/Partitionin	Sun Solaris 10 10/09	Tuxedo CFS-R	11/03/09	Y
3	IBM	IBM Power 595 Server Model 9119-FHA	6,085,166	2.81 USD	NR	12/10/08	IBM DB2 9.5	IBM AIX 5L V5.3	Microsoft COM+	06/10/08	N
***	BULL	Bull Escala PL6460R	6,085,166	2.81 USD	NR	12/15/08	IBM DB2 9.5	IBM AIX 5L V5.3	Microsoft COM+	06/15/08	N
4	hp invent	HP Integrity Superdome-Itanium2/1.6GHz/24MB iL3	4,092,799	2.93 USD	NR	08/06/07	Oracle Database 10g R2 Enterprise Edt w/Partitioning	HP-UX 11i v3	BEA Tuxedo 8.0	02/27/07	N
5	IBM	IBM System p5 595	4,033,378	2.97 USD	NR	01/22/07	IBM DB2 9	IBM AIX 5L V5.3	Microsoft COM+	01/22/07	N
6	IBM	IBM eServer p5 595	3,210,540	5.07 USD	NR	05/14/05	IBM DB2 UDB 8.2	IBM AIX 5L V5.3	Microsoft COM+	11/18/04	N
7	FUJITSU	PRIMEQUEST 580A 32p/64c	2,382,032	3.76 USD	NR	12/04/08	Oracle Database 10g R2 Enterprise Edt w/Partitioning	Red Hat Enterprise Linux 4 AS	BEA Tuxedo 8.1	12/04/08	N

RAS



POWER7 – funkcje niezawodnościowe (RAS)

Operating System
Hot patch Kernel
Storage Keys

Memory
Chip Kill technology with Bit-steering
Selective Mirroring

Hot Plug / Removal
Fans & Power Supplies

Hot Plug / Removal
PCI-X Adapters
IO Drawers

Hot Plug / Removal
Disks

Hot Add
I/O racks

Processors
Dynamic De-Allocation
Packaging
Instruction Retry
Alternate Processor Recovery

Mobility
Partition Mobility
WPAR Mobility

Dual Clocks
570 / 595 / 770 / 780 / 795

Hot Add (570, 595, 770, 780, 795)
Eliminates Upgrade outages

Concurrent Service (570, 595, 770, 780, 795)
Eliminates Repair Outages

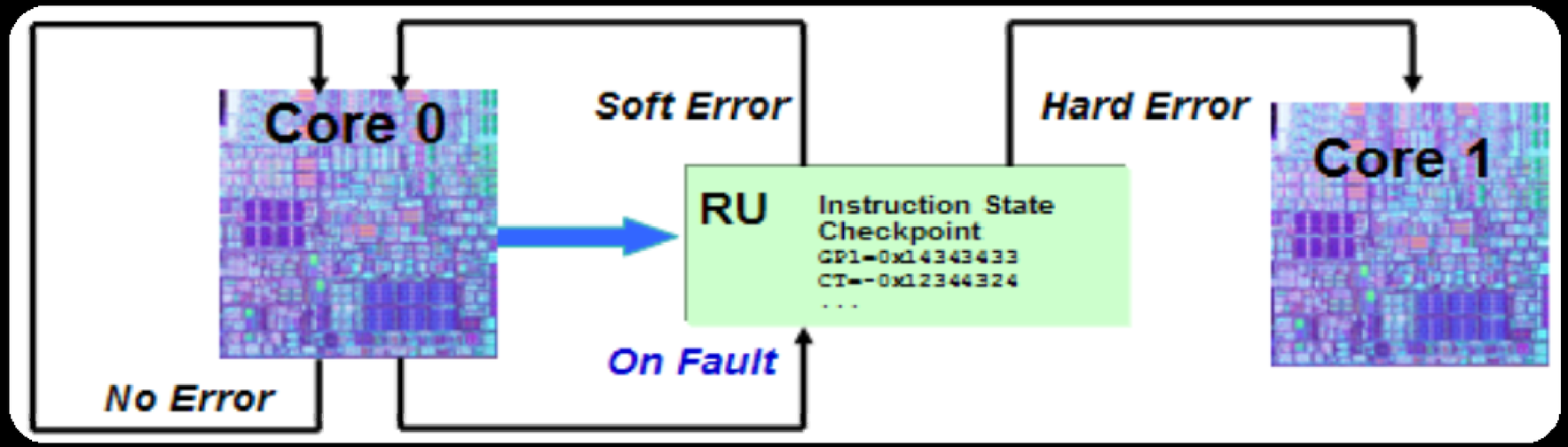
Hypervisor
Mainframe technology

Passive backplane
No active components

First Failure Data Capture
Help eliminates intermittent failures

Alternate Processor Recovery

- Soft Error
  - stany procesora są zapamiętywane i zabezpieczane za pomocą ECC
  - w razie wystąpienia błędu instrukcje mogą być powtórzone
- Hard Error
  - jeśli operacja ponownie generuje błąd, stan procesora (rdzenia) jest przenoszony na inny rdzeń w sposób całkowicie przezroczysty dla systemu operacyjnego
  - uszkodzony rdzeń jest dealokowany
  - jeśli serwer posiada rdzenie nieaktywne (CUoD), zostaną automatyczne uruchomione

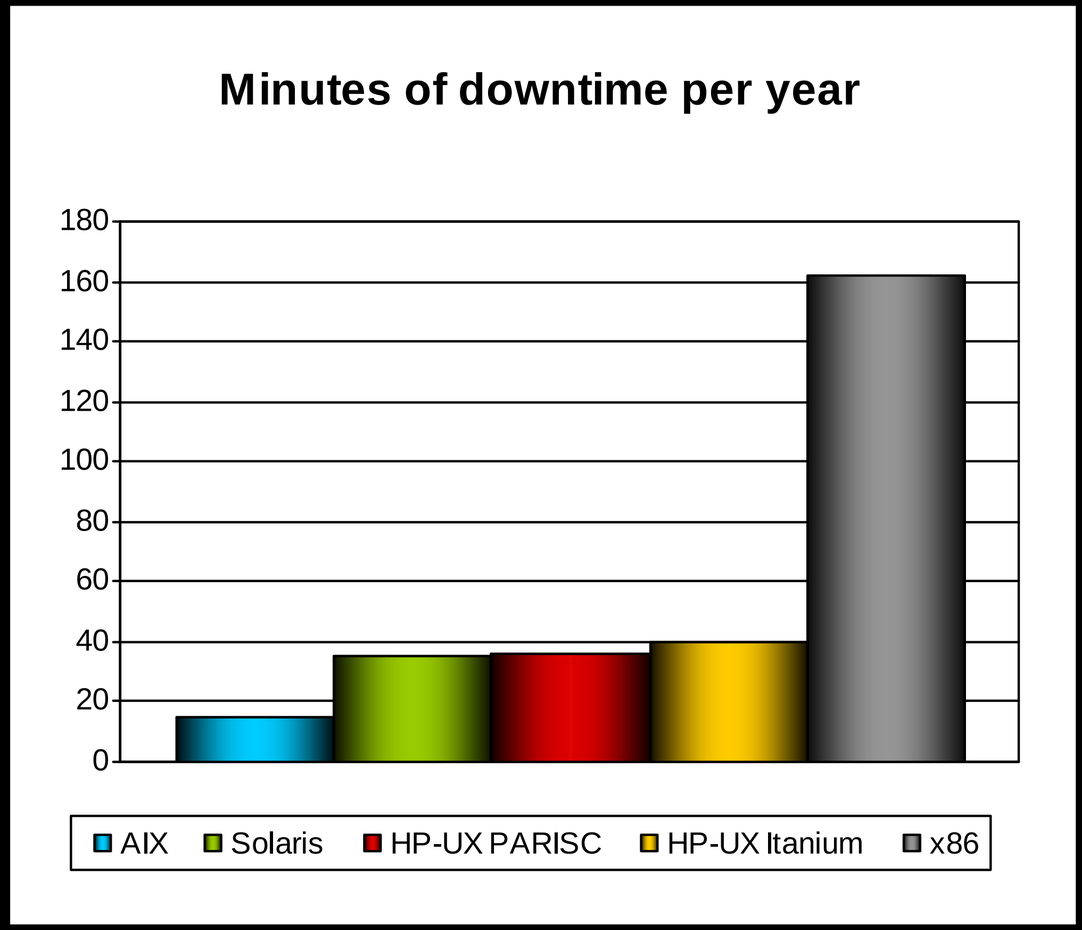


Funkcje RAS – porównanie

RAS Feature	POWER	SPARC	Integrity	Xeon
Application/Partition RAS				
Live Partition Mobility	Yes	No	No	Yes
Live Application Mobility	Yes	No	No	No
Partition Availability priority	Yes	No	No	No
System RAS				
OS independent First Failure Data Capture	Yes	No	No	No
Redundant System Interconnect	No	Yes	No	No
Memory Keys	Yes	No	No	No
Processor RAS				
Processor Instruction Retry	Yes	Yes	No	No
Alternate Processor Recovery	Yes	No	No	No
Dynamic Processor Deallocation	Yes	Yes	Yes	No
Dynamic Processor Sparing	Yes	Partial	Partial	No
Memory RAS				
Chipkill™	Yes	Yes	Yes	Yes
Redundant Memory	Yes	Yes	Yes	Yes
I/O RAS				
Extended Error Handling	Yes	No	No	No

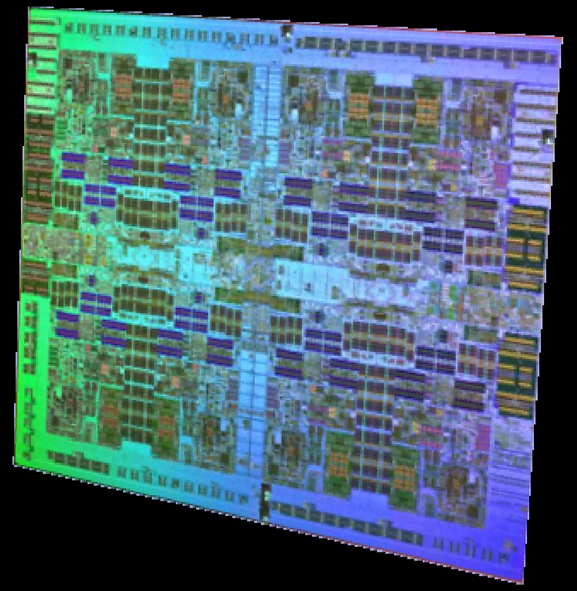
Niezawodność systemów – ITIC 2009

- Systemy Power z systemem AIX oferują najbardziej zaawansowany RAS wśród systemów otwartych
- Najkrótsze czasy przestoju
  – średnio 15 minut w roku
- Ponad 2x lepsza dostępność niż najbliższy konkurent z rodziny UNIX
- Ponad 10x lepsza niezawodność niż serwer z systemem Windows
- Najmniej nieplanowanych przestojów (średnio < 1 w roku)
- Najkrótszy czas instalowania poprawek (średnio 11 minut)

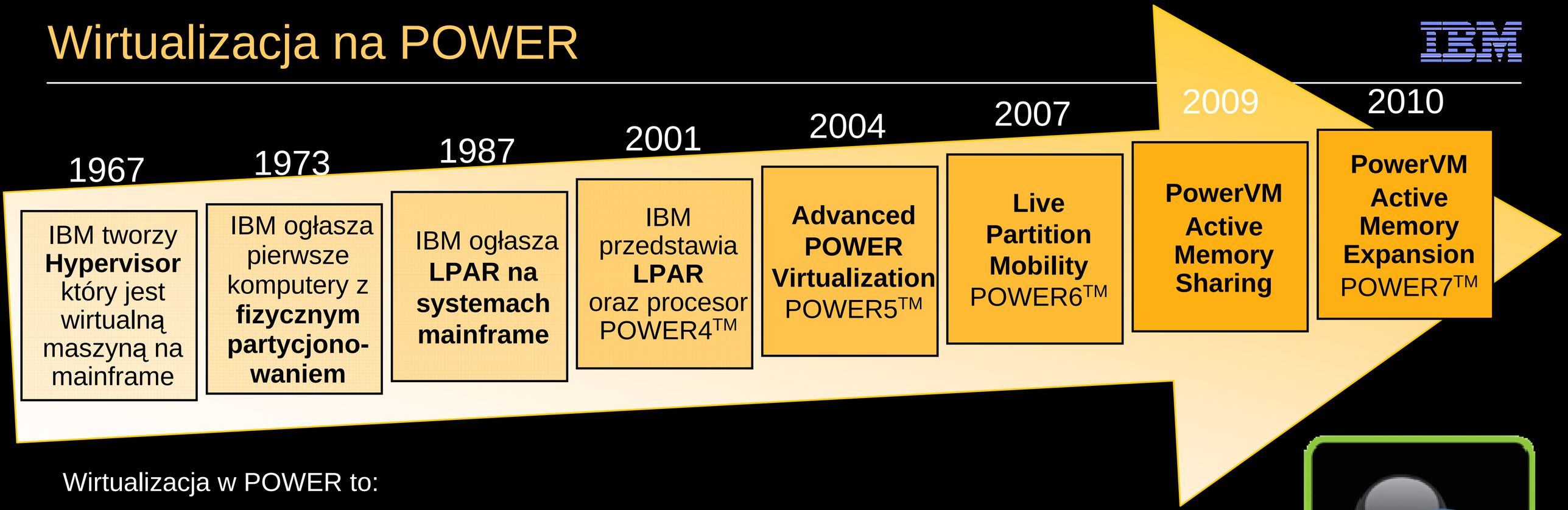


http://itic-corp.com/blog/2009/07/itic-2009-global-server-hardware-server-os-reliability-survey-results/

PowerVM



Wirtualizacja na POWER

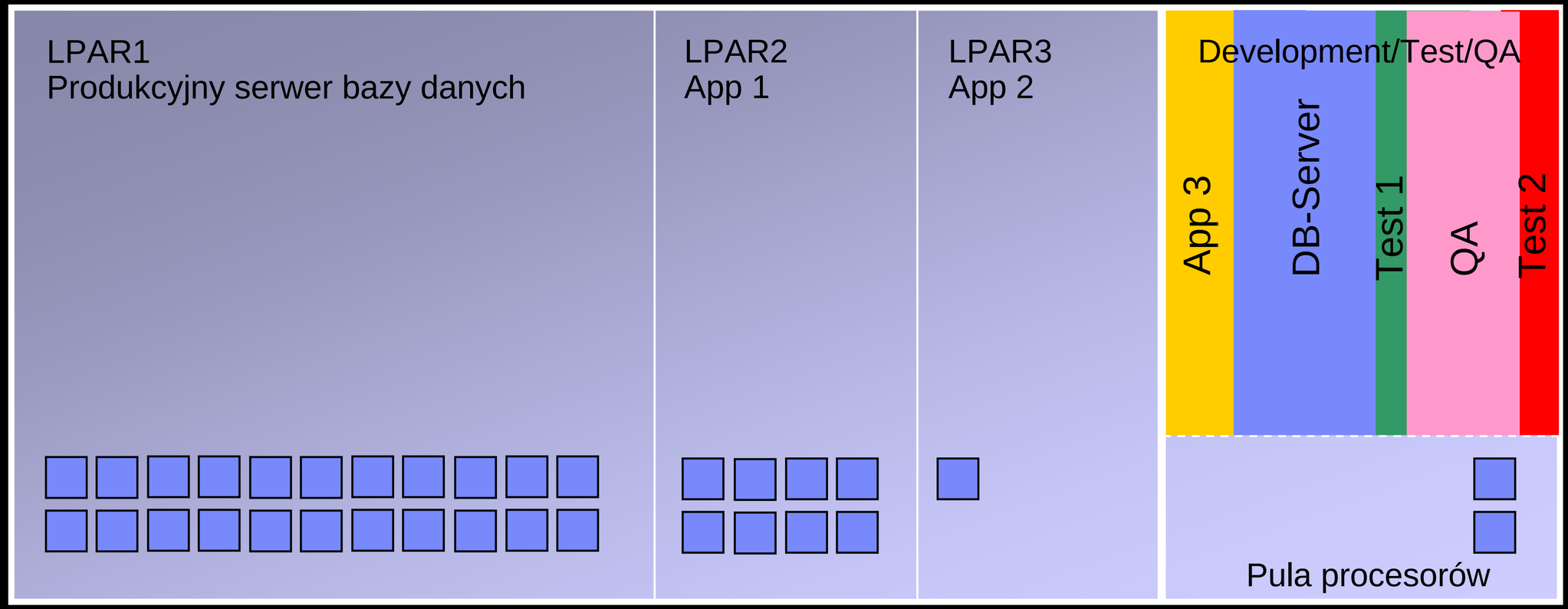


Wirtualizacja w POWER to:

- hypervisor zaprojektowany i zintegrowany ze sprzętem, z minimalnym narzutem na wydajność,
- izolacja zasobów gwarantowana sprzętowo i potwierdzona certyfikatami bezpieczeństwa,
- precyzyjny i w pełni dynamiczny przydział zasobów CPU, RAM, I/O,
- możliwość dedykowania dowolnych zasobów lub współdzielenia cykli procesorów dedykowanych,
- ekstremalna skalowalność i niezawodność, redundancja VIOS,
- moc na żądanie (aktywacja zasobów uśpionych – CPU i RAM),
- przenoszenie maszyn wirtualnych w czasie ich pracy między fizycznymi serwerami,
- kompresja pamięci RAM w sposób przezroczysty dla aplikacji.

PowerVM™

PowerVM – partycjonowanie w pełni dynamiczne



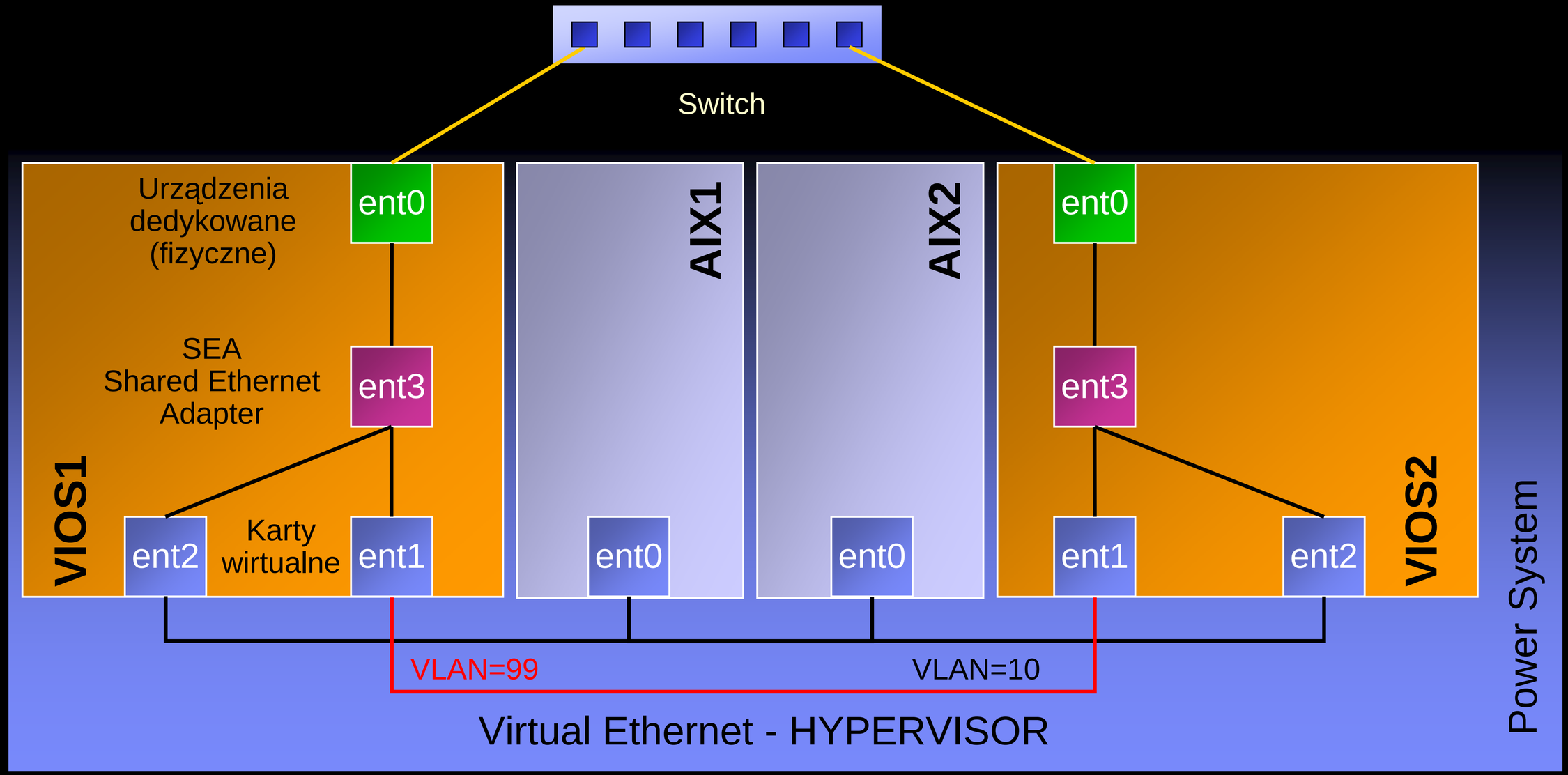
Dynamiczna relokacja dedykowanych zasobów do wybranej puli procesorów

Dynamiczna relokacja zasobów między partycjami dedykowanymi

Ciągły load-balancing w ramach puli procesorów

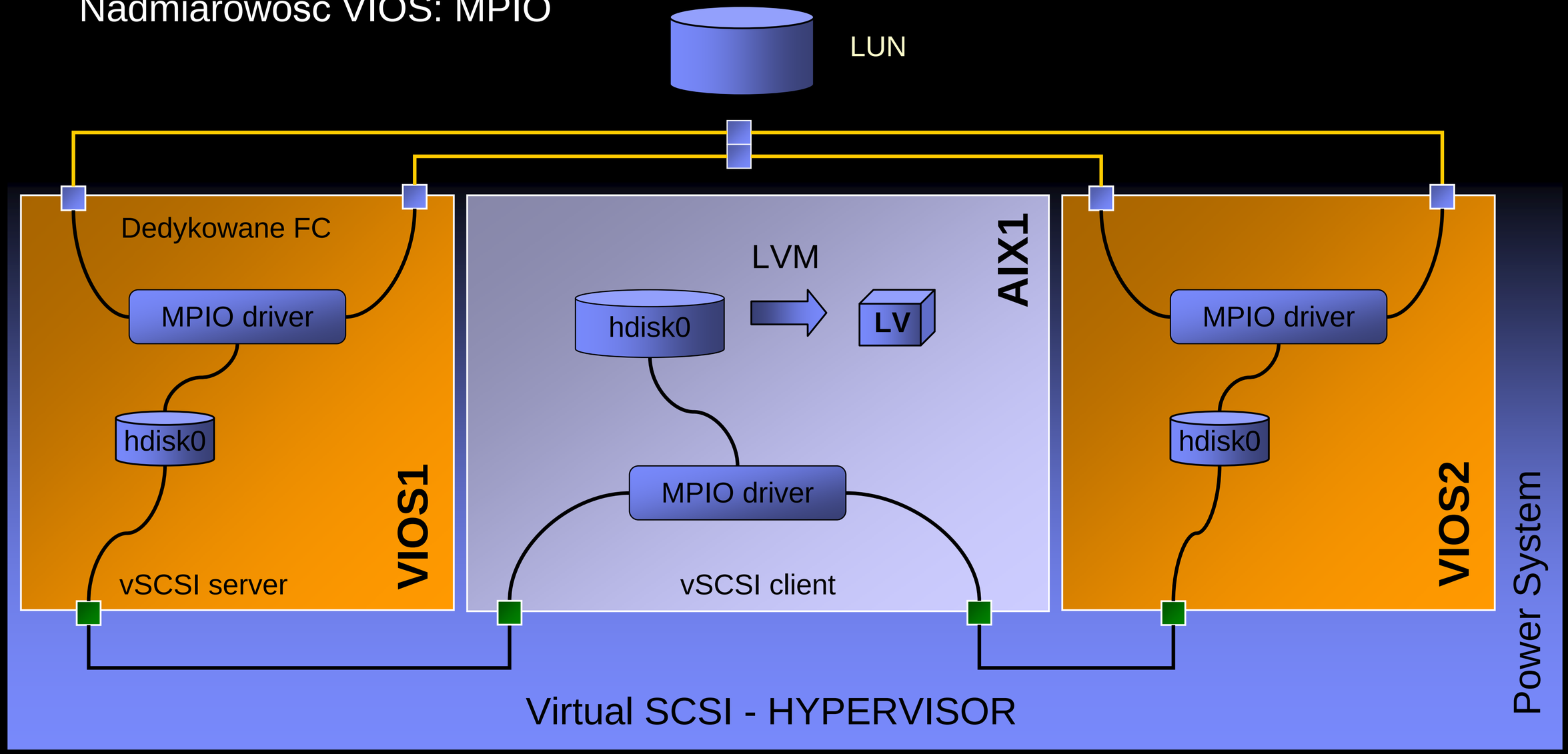
PowerVM – wysoka dostępność VIOS

Nadmiarowość VIOS: Shared Ethernet Adapter Failover



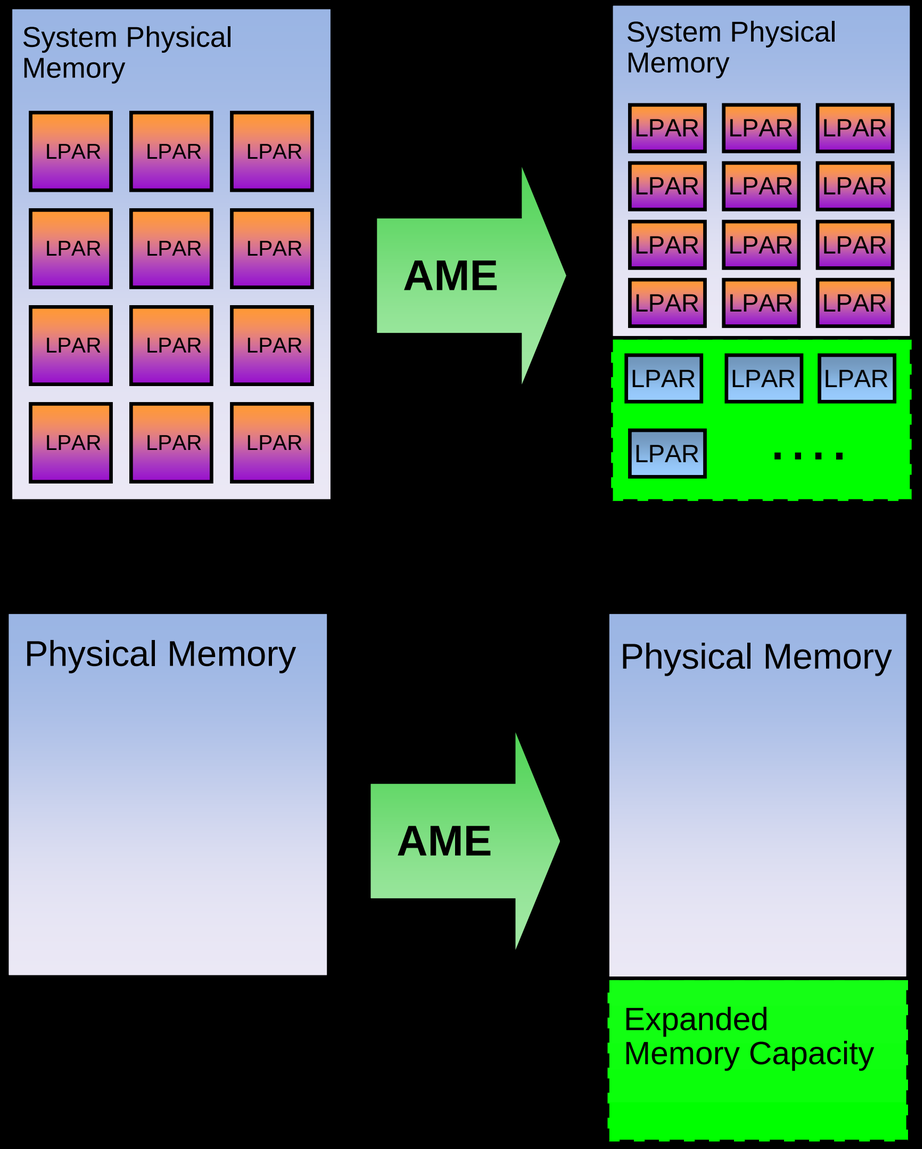
PowerVM – wysoka dostępność VIOS

Nadmiarowość VIOS: MPIO

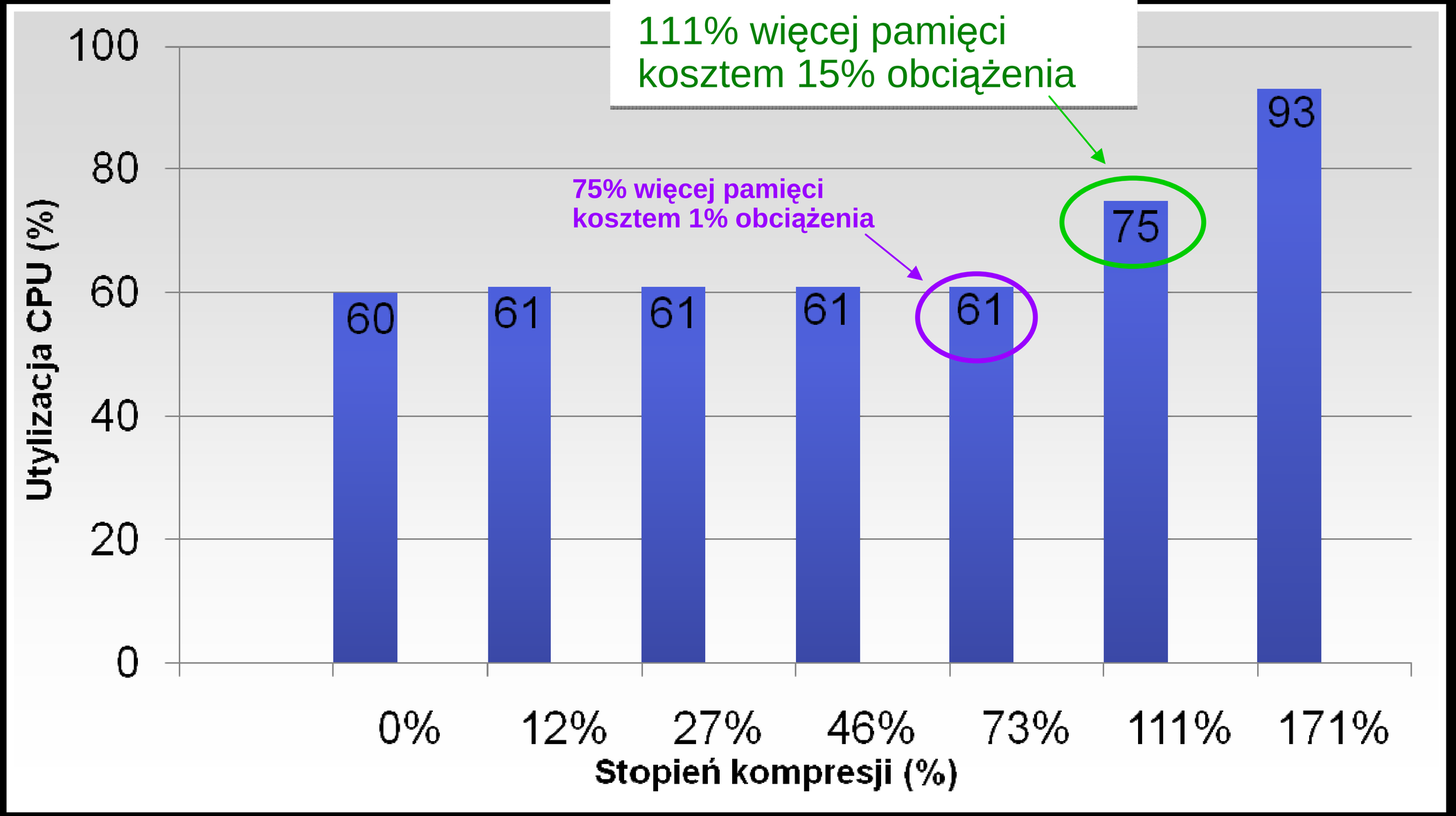


PowerVM – Active Memory Expansion

- AME kompresuje strony pamięci używając mocy obliczeniowej procesorów
- Kompresja jest przezroczysta dla aplikacji – wykonuje ją hypervisor
- Konfigurowalna niezależnie dla każdej partycji
- Możliwe dynamiczne uruchomienie lub zmiana stopnia kompresji
- Stopień kompresji oraz zapotrzebowanie na moc obliczeniową jest zależne od uruchomionej aplikacji



Active Memory Expansion w SAP ERP



Dziękuję za uwagę