MAKLEE

software engineering solutions

OpenVMS on BL890c i2 Servers

Guy Peleg
President
Maklee Engineering
guy.peleg@maklee.com

Maklee Engineering

ORACLE Gold Partner

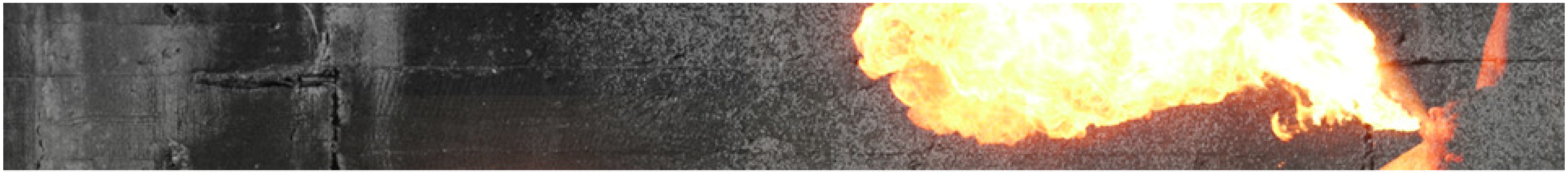
- Consulting firm operating all over the world.
  - Team of “Top Gun” engineers.
  - Former members of various engineering groups at HP.
  - Gold Oracle partner.

- Specialize in performance tuning of:
  - OpenVMS
  - Oracle (HP-UX, Linux, VMS, Solaris, AIX, Windows)
  - Oracle Rdb
  - Java (HP-UX, Linux, VMS, Solaris, AIX, Windows)
  - Adabas

- Also offers custom engineering services, custom training and on-going support contracts.

MAKLEE

Maklee provides guarantee for success for all projects !!

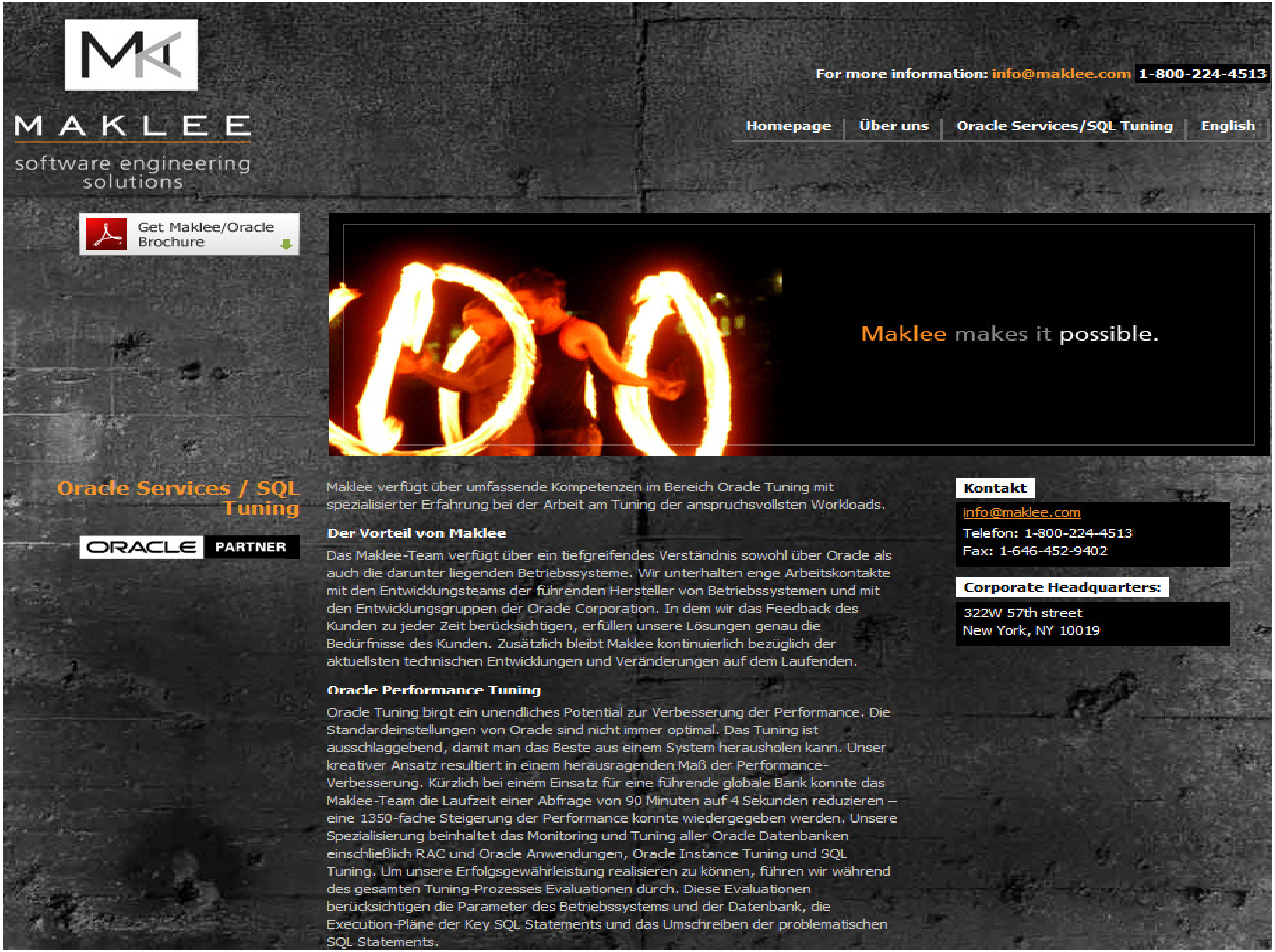


Maklee Guarantees <u>doubling</u> the performance of your Oracle database or our service is provided free of charge !

and….we speak German !!

http://www.maklee.com/indexDeutsch.html

MAKLEE


MAKLEE
software engineering solutions
Homepage
Über uns
Oracle Services/SQL Tuning
English
Get Maklee/Oracle Brochure
ORACLE PARTNER

For more information: info@maklee.com 1-800-224-4513

Maklee makes it possible.

Oracle Services / SQL Tuning

Maklee verfügt über umfassende Kompetenzen im Bereich Oracle Tuning mit spezialisierter Erfahrung bei der Arbeit am Tuning der anspruchsvollsten Workloads.

Der Vorteil von Maklee

Das Maklee-Team verfügt über ein tiefgreifendes Verständnis sowohl über Oracle als auch die darunter liegenden Betriebssysteme. Wir unterhalten enge Arbeitskontakte mit den Entwicklungsteams der führenden Hersteller von Betriebssystemen und mit den Entwicklungsgruppen der Oracle Corporation. In dem wir das Feedback des Kunden zu jeder Zeit berücksichtigen, erfüllen unsere Lösungen genau die Bedürfnisse des Kunden. Zusätzlich bleibt Maklee kontinuierlich bezüglich der aktuellsten technischen Entwicklungen und Veränderungen auf dem Laufenden.

Oracle Performance Tuning

Oracle Tuning birgt ein unendliches Potential zur Verbesserung der Performance. Die Standardeinstellungen von Oracle sind nicht immer optimal. Das Tuning ist ausschlaggebend, damit man das Beste aus einem System herausholen kann. Unser kreativer Ansatz resultiert in einem herausragenden Maß der Performance-Verbesserung. Kürzlich bei einem Einsatz für eine führende globale Bank konnte das Maklee-Team die Laufzeit einer Abfrage von 90 Minuten auf 4 Sekunden reduzieren – eine 1350-fache Steigerung der Performance konnte wiedergegeben werden. Unsere Spezialisierung beinhaltet das Monitoring und Tuning aller Oracle Datenbanken einschließlich RAC und Oracle Anwendungen, Oracle Instance Tuning und SQL Tuning. Um unsere Erfolgsgewährleistung realisieren zu können, führen wir während des gesamten Tuning-Prozesses Evaluationen durch. Diese Evaluationen berücksichtigen die Parameter des Betriebssystems und der Datenbank, die Execution-Pläne der Key SQL Statements und das Umschreiben der problematischen SQL Statements.

Kontakt

info@maklee.com
Telefon: 1-800-224-4513
Fax: 1-646-452-9402

Corporate Headquarters:

322W 57th street
New York, NY 10019

BL890c i2

- Why do we need to spend the next hour discussing OpenVMS on the new BL890c i2 server?

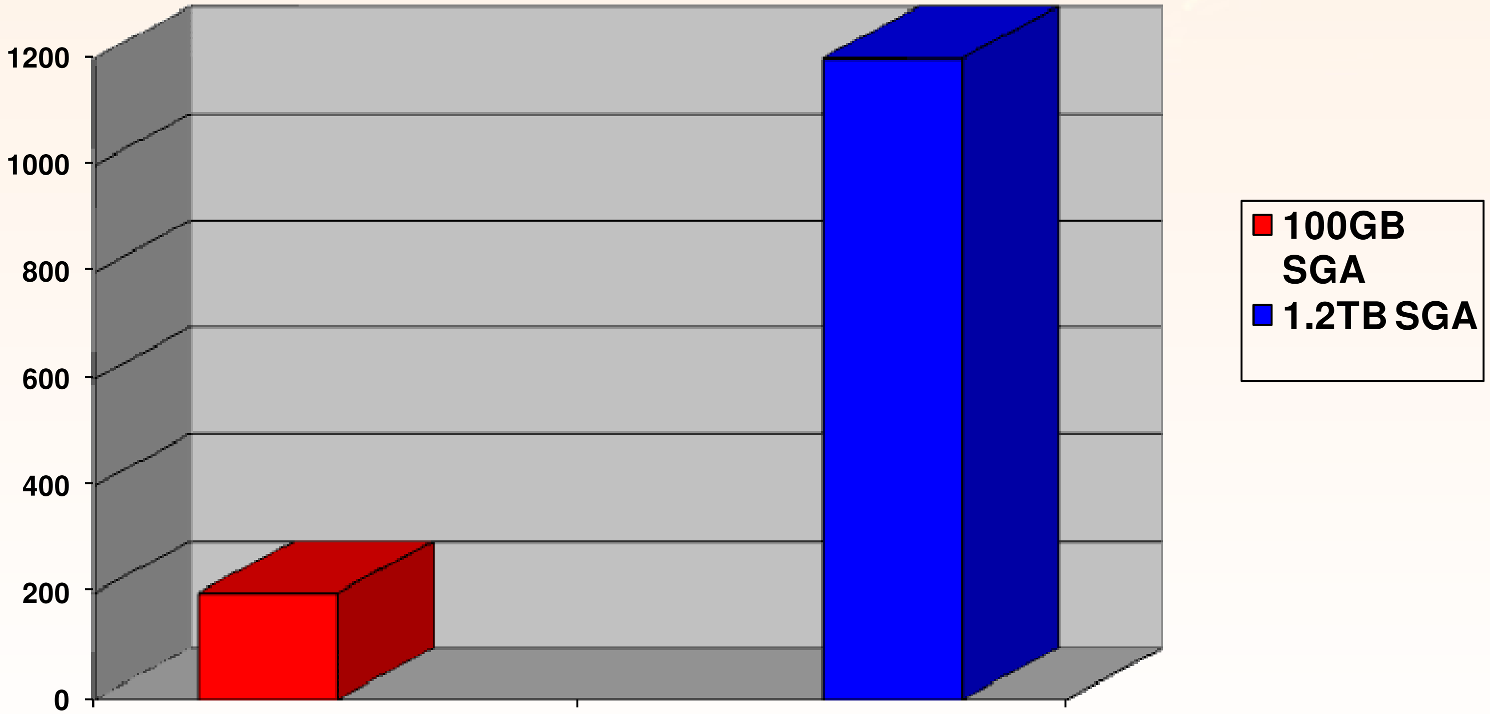
- What's unique about the new server?

- What happened to VMS is VMS is VMS ?

- The BL890c i2 was built using a new memory architecture.
  - Understanding the new architecture is essential for achieving optimal performance on the new server.

Extreme Example

- Superdome2 , 2TB RAM, HP-UX 11.31 (update 7)
- Oracle 11gR2



Elapsed time (ms) to execute single select query

Memory Latency and NUMA

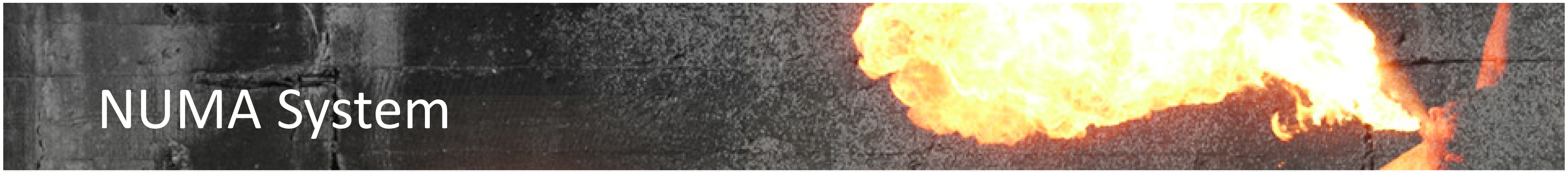
- The CPU is MUCH faster than physical memory.
  - CPU cycle is ~0.5 nanosecond.

- Memory latency is the amount of time it takes for data to arrive from physical memory into the CPU.
  - Varies from 40 – 500ns
  - 80-1000 times slower than the CPU

- Most CPUs spend significant amount of time waiting for data to arrive from physical memory.
  - From VMS perspective the CPU looks busy

- On a Non Uniform Memory Access architecture (NUMA) accessing local memory is faster than remote memory.

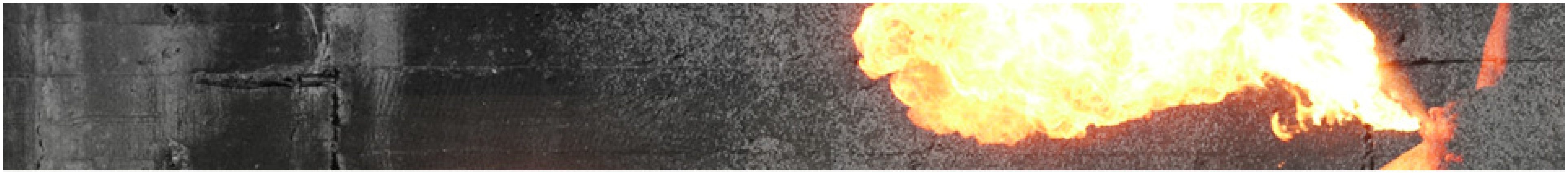
MAKLEE



NUMA System

Building Block #0 CPUs 0-7 Memory (interleaved) SCHED and SCS spinlock Disk and Network I/O adapters	Building Block #1 CPUs 8-15 Memory (interleaved) Disk and Network I/O adapters
Building Block #2 CPUs 16-23 Memory (interleaved) LCKMGR and TCPIP spinlock	Building Block #3 CPUs 24-31 Memory (interleaved)

Life is not fair !!



OKAY !

BUT….does it really matter??

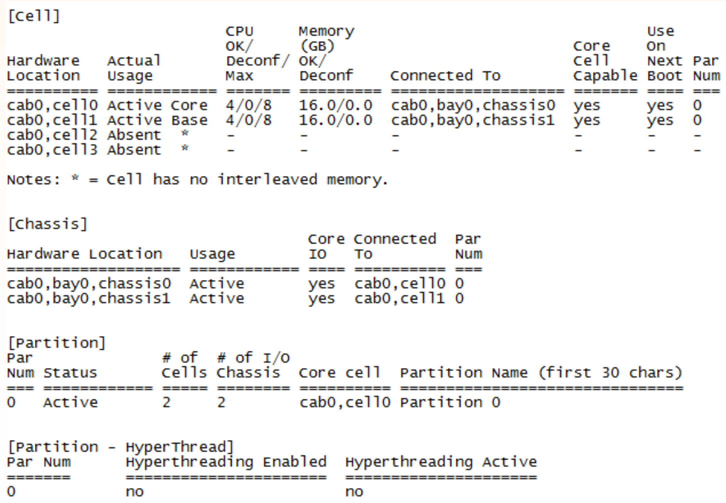
Oh YES !!!

MAKLEE

Memory Latency

- 2 Cells 4P/8C rx8640 Integrity server.
- In preparation to future growth, a customer purchased 4 processors and spread them across 2 cells.
- 32GB RAM.
- Noticed very high CPU utilization comparing to older integrity box running the same workload.
- Maklee recommended consolidating all of the processors into a single cell, power off the second cell, and by that improve memory latency.

CPU UTILIZATION DROPPED 40%



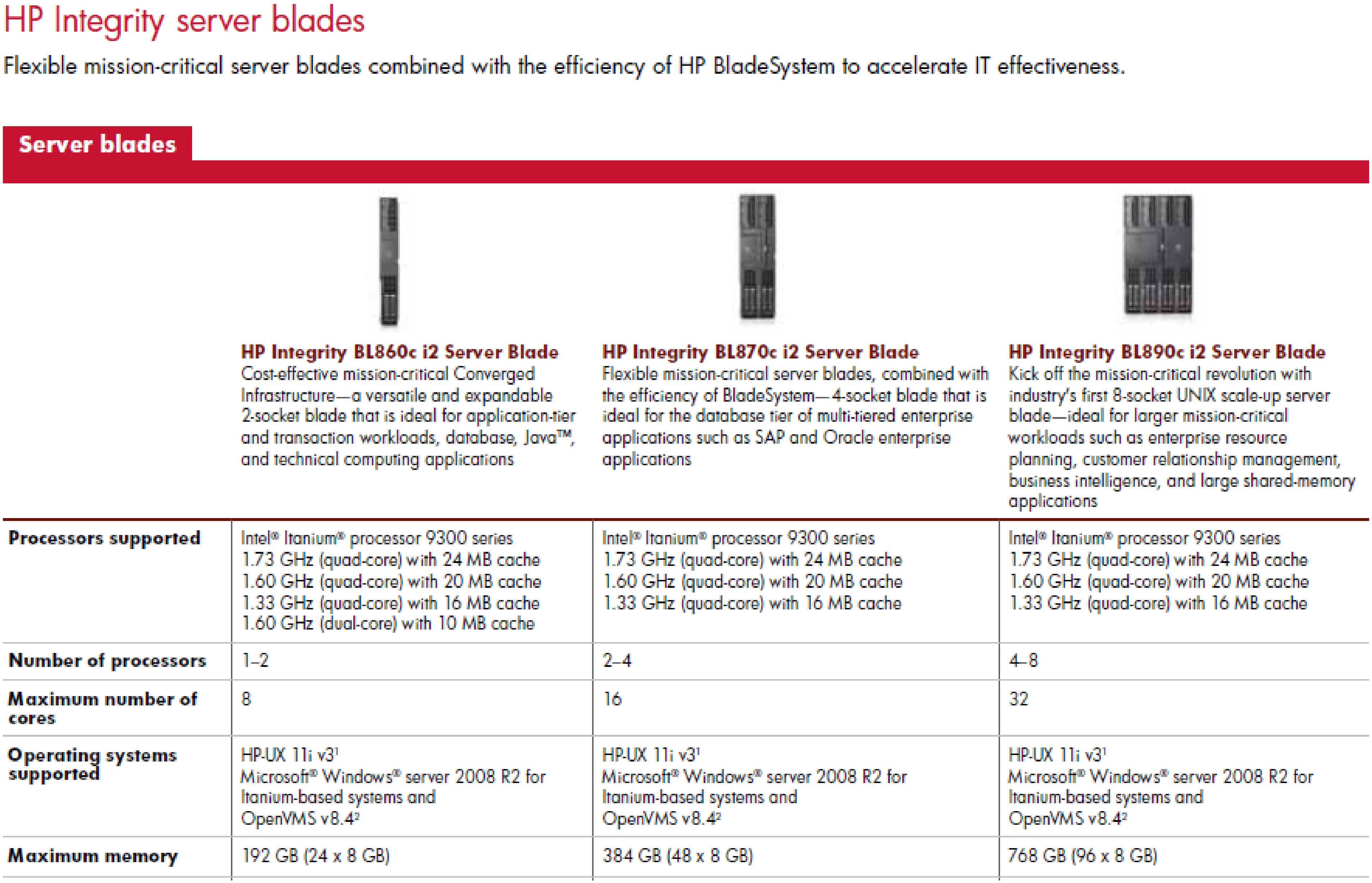
```
[Cell]
                        CPU     Memory                                      Use
                        OK/     (GB)                          Core    On
Hardware   Actual       Deconf/ OK/                           Cell    Next Par
Location   Usage        Max     Deconf    Connected To        Capable Boot Num
========== ============ ======= ========= =================== ======= ==== ===
cab0,cell0 Active Core  4/0/8   16.0/0.0  cab0,bay0,chassis0  yes     yes  0
cab0,cell1 Active Base  4/0/8   16.0/0.0  cab0,bay0,chassis1  yes     yes  0
cab0,cell2 Absent   *   -       -         -                   -       -    -
cab0,cell3 Absent   *   -       -         -                   -       -    -

Notes: * = Cell has no interleaved memory.


[Chassis]
                                 Core Connected  Par
Hardware Location   Usage        IO   To         Num
=================== ============ ==== ========== ===
cab0,bay0,chassis0  Active       yes  cab0,cell0 0
cab0,bay0,chassis1  Active       yes  cab0,cell1 0


[Partition]
Par              # of  # of I/O
Num Status       Cells Chassis  Core cell  Partition Name (first 30 chars)
=== ============ ===== ======== ========== ===============================
0   Active       2     2        cab0,cell0 Partition 0


[Partition - HyperThread]
Par Num      Hyperthreading Enabled  Hyperthreading Active
=======      ======================  =====================
0            no                      no
```

The Golden Rules

- Run your application on the smallest Integrity server that fits your workload.
  - rx6600 and blade BL870c provide outstanding performance, with very low memory latency.
  - On a cellular system, do not turn on extra cells unless you REALLY need it.
- For workloads that do not fit a small system:

A process should be "close" to it's memory

New Line of Integrity Servers



HP Integrity server blades

Flexible mission-critical server blades combined with the efficiency of HP BladeSystem to accelerate IT effectiveness.

Server blades

	HP Integrity BL860c i2 Server Blade Cost-effective mission-critical Converged Infrastructure—a versatile and expandable 2-socket blade that is ideal for application-tier and transaction workloads, database, Java™, and technical computing applications	**HP Integrity BL870c i2 Server Blade** Flexible mission-critical server blades, combined with the efficiency of BladeSystem—4-socket blade that is ideal for the database tier of multi-tiered enterprise applications such as SAP and Oracle enterprise applications	**HP Integrity BL890c i2 Server Blade** Kick off the mission-critical revolution with industry's first 8-socket UNIX scale-up server blade—ideal for larger mission-critical workloads such as enterprise resource planning, customer relationship management, business intelligence, and large shared-memory applications
Processors supported	Intel® Itanium® processor 9300 series 1.73 GHz (quad-core) with 24 MB cache 1.60 GHz (quad-core) with 20 MB cache 1.33 GHz (quad-core) with 16 MB cache 1.60 GHz (dual-core) with 10 MB cache	Intel® Itanium® processor 9300 series 1.73 GHz (quad-core) with 24 MB cache 1.60 GHz (quad-core) with 20 MB cache 1.33 GHz (quad-core) with 16 MB cache	Intel® Itanium® processor 9300 series 1.73 GHz (quad-core) with 24 MB cache 1.60 GHz (quad-core) with 20 MB cache 1.33 GHz (quad-core) with 16 MB cache
Number of processors	1–2	2–4	4–8
Maximum number of cores	8	16	32
Operating systems supported	HP-UX 11i v3[1] Microsoft® Windows® server 2008 R2 for Itanium-based systems and OpenVMS v8.4[2]	HP-UX 11i v3[1] Microsoft® Windows® server 2008 R2 for Itanium-based systems and OpenVMS v8.4[2]	HP-UX 11i v3[1] Microsoft® Windows® server 2008 R2 for Itanium-based systems and OpenVMS v8.4[2]
Maximum memory	192 GB (24 x 8 GB)	384 GB (48 x 8 GB)	768 GB (96 x 8 GB)

BL890c i2

MAKLEE

The Tukwila Processor

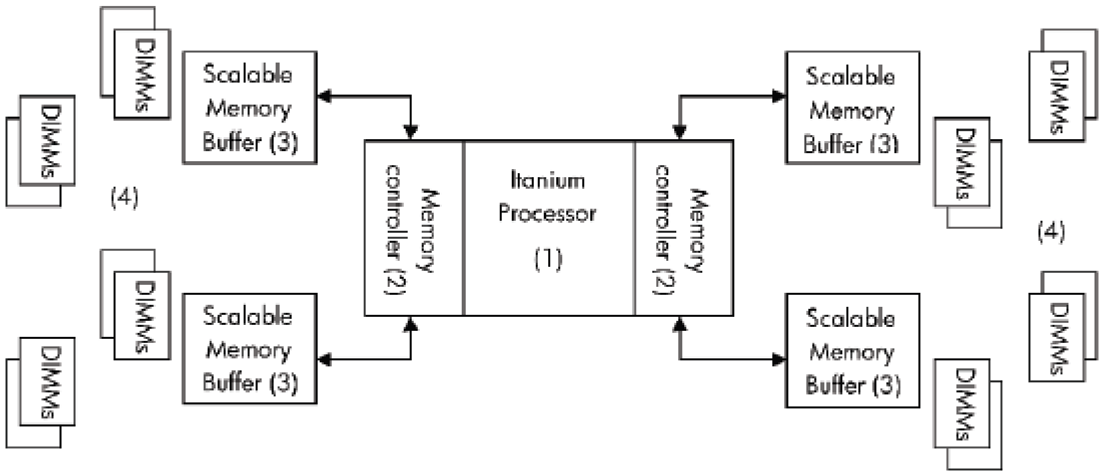
- **Tukwila** is the code-name for the generation of Intel's Itanium processor family following Itanium 2 , Montecito and Montvale. It was released on 8 February 2010 as the Itanium 9300 Series.

- Quad Core processor, 1.73GHz, 6MB L3 cache per core.

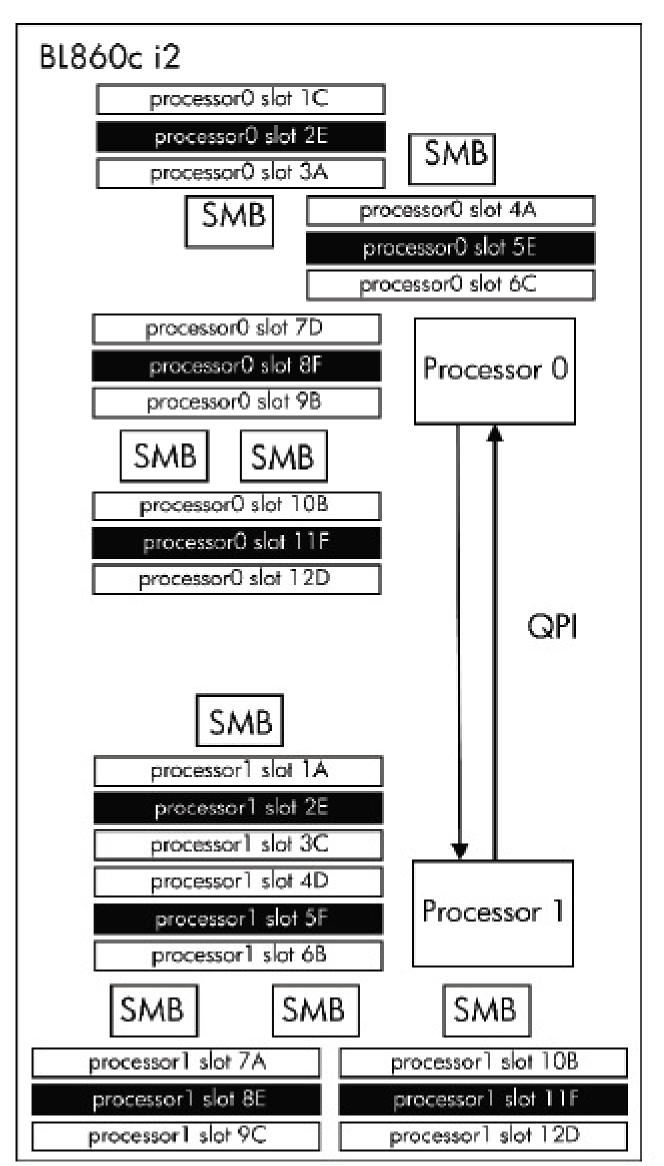
- Socket compatibility between Intel's Xeon and Itanium processors, by introducing a new interconnect called Intel QuickPath Interconnect (QPI).
  - Point-to-Point processor interconnect.
  - Allows one processor module to access memory connected to other processor module.
  - Developed by members of what had been DEC's Alpha group.
  - Replaces the Front Side Bus (FSB) for Xeon and Itanium.
  - First delivered on the Intel Core i7-9xx desktop processors and the X58 chipset.
  - The memory controller is part of the processor module and not the chipset.

Memory Subsystem Overview

Figure 1: High-level Itanium Processor Memory Subsystem Diagram

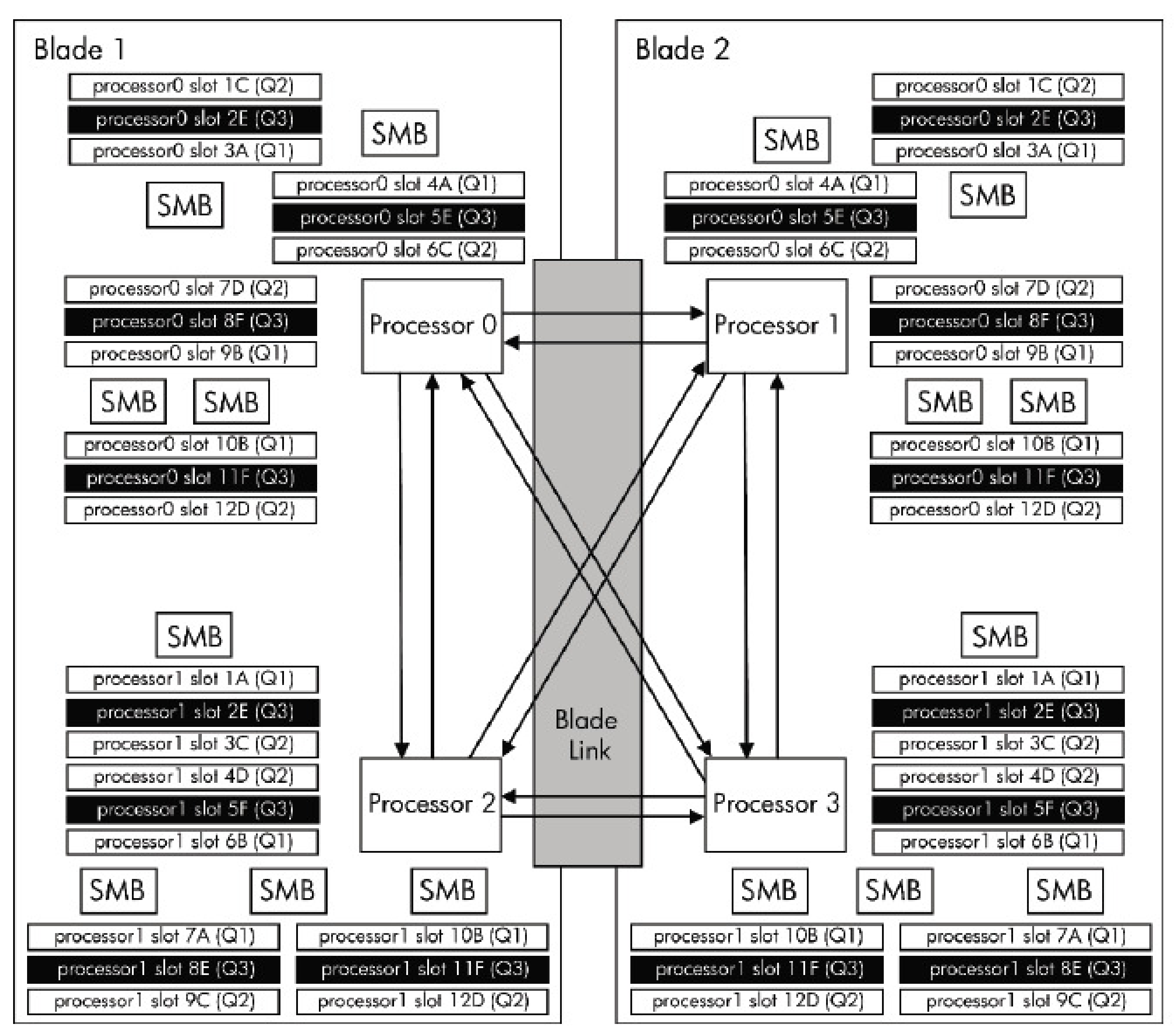


BL860c i2 Overview



MAKLEE

BL870c i2 Overview



Blade 1

processor0 slot 1C (Q2)
processor0 slot 2E (Q3)
processor0 slot 3A (Q1)

SMB

SMB

processor0 slot 4A (Q1)
processor0 slot 5E (Q3)
processor0 slot 6C (Q2)

processor0 slot 7D (Q2)
processor0 slot 8F (Q3)
processor0 slot 9B (Q1)

SMB SMB

processor0 slot 10B (Q1)
processor0 slot 11F (Q3)
processor0 slot 12D (Q2)

Processor 0

SMB

processor1 slot 1A (Q1)
processor1 slot 2E (Q3)
processor1 slot 3C (Q2)
processor1 slot 4D (Q2)
processor1 slot 5F (Q3)
processor1 slot 6B (Q1)

Processor 2

SMB SMB SMB

processor1 slot 7A (Q1)
processor1 slot 8E (Q3)
processor1 slot 9C (Q2)

processor1 slot 10B (Q1)
processor1 slot 11F (Q3)
processor1 slot 12D (Q2)

Blade Link

Blade 2

processor0 slot 1C (Q2)
processor0 slot 2E (Q3)
processor0 slot 3A (Q1)

SMB

processor0 slot 4A (Q1)
processor0 slot 5E (Q3)
processor0 slot 6C (Q2)

SMB

Processor 1

processor0 slot 7D (Q2)
processor0 slot 8F (Q3)
processor0 slot 9B (Q1)

SMB SMB

processor0 slot 10B (Q1)
processor0 slot 11F (Q3)
processor0 slot 12D (Q2)

SMB

processor1 slot 1A (Q1)
processor1 slot 2E (Q3)
processor1 slot 3C (Q2)
processor1 slot 4D (Q2)
processor1 slot 5F (Q3)
processor1 slot 6B (Q1)

Processor 3

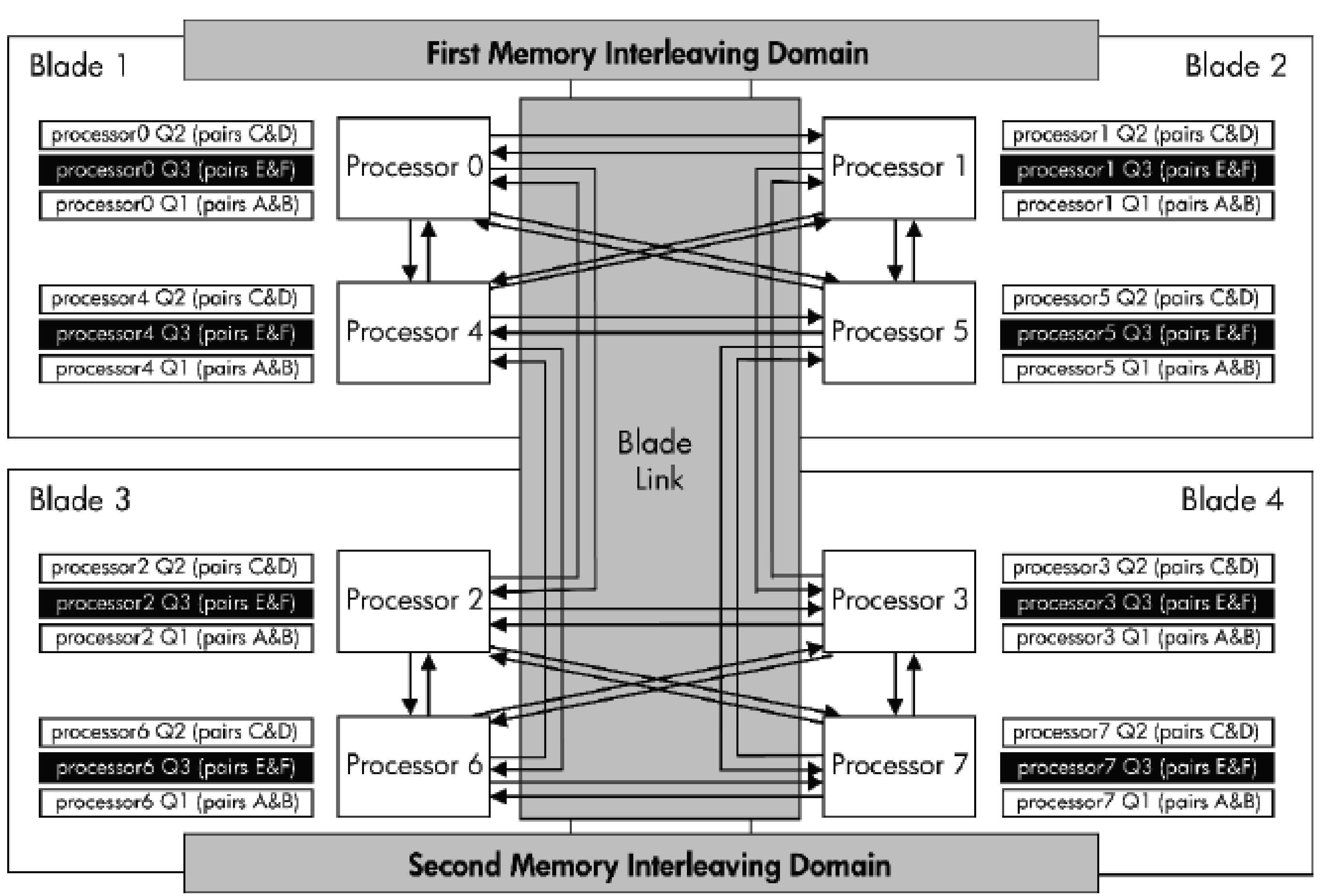
SMB SMB SMB

processor1 slot 10B (Q1)
processor1 slot 11F (Q3)
processor1 slot 12D (Q2)

processor1 slot 7A (Q1)
processor1 slot 8E (Q3)
processor1 slot 9C (Q2)

MAKLEE

BL890c i2 Overview



First Memory Interleaving Domain

Blade 1

processor0 Q2 (pairs C&D)
processor0 Q3 (pairs E&F)
processor0 Q1 (pairs A&B)

Processor 0

processor4 Q2 (pairs C&D)
processor4 Q3 (pairs E&F)
processor4 Q1 (pairs A&B)

Processor 4

Blade 2

Processor 1

processor1 Q2 (pairs C&D)
processor1 Q3 (pairs E&F)
processor1 Q1 (pairs A&B)

Processor 5

processor5 Q2 (pairs C&D)
processor5 Q3 (pairs E&F)
processor5 Q1 (pairs A&B)

Blade Link

Blade 3

processor2 Q2 (pairs C&D)
processor2 Q3 (pairs E&F)
processor2 Q1 (pairs A&B)

Processor 2

processor6 Q2 (pairs C&D)
processor6 Q3 (pairs E&F)
processor6 Q1 (pairs A&B)

Processor 6

Blade 4

Processor 3

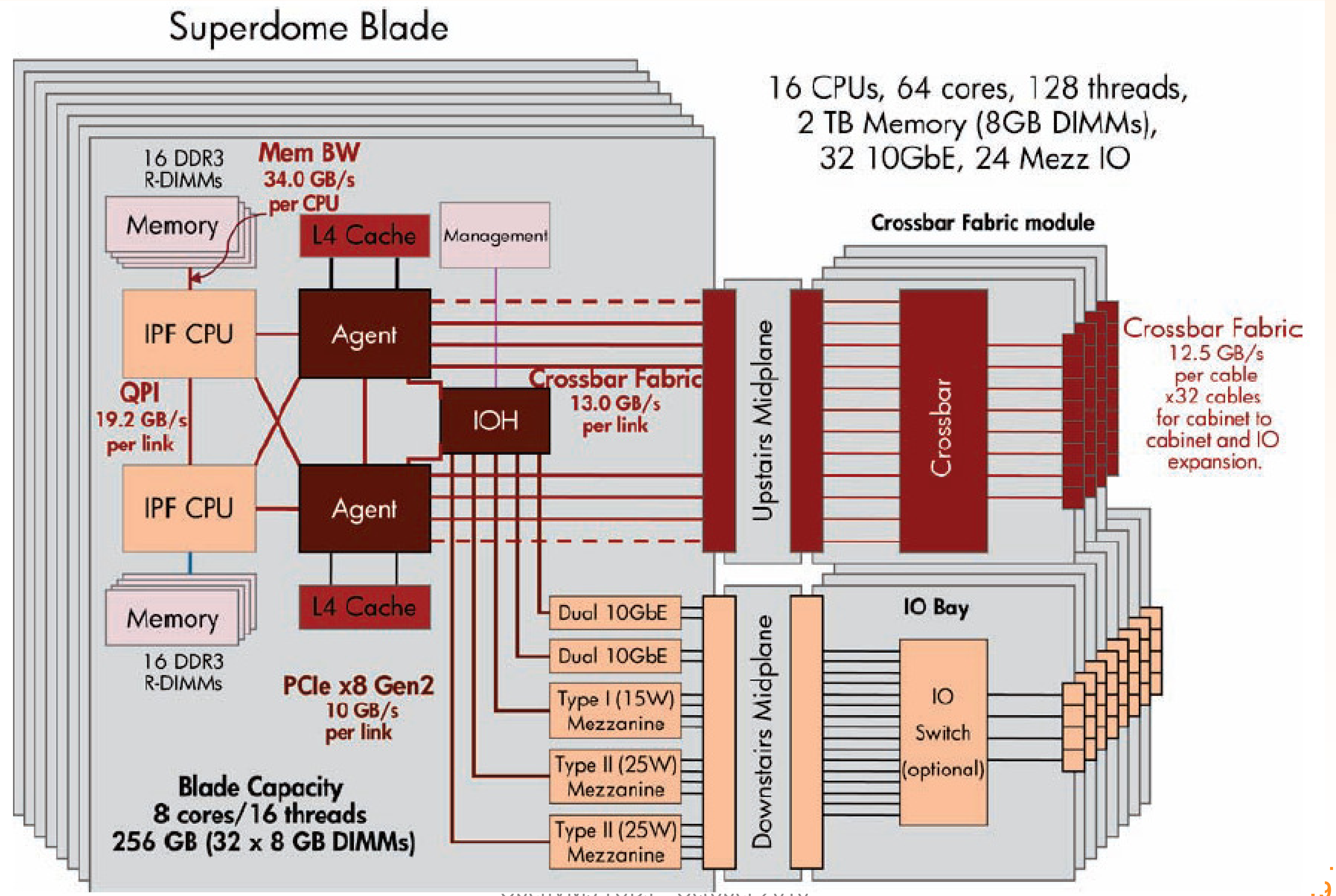
processor3 Q2 (pairs C&D)
processor3 Q3 (pairs E&F)
processor3 Q1 (pairs A&B)

Processor 7

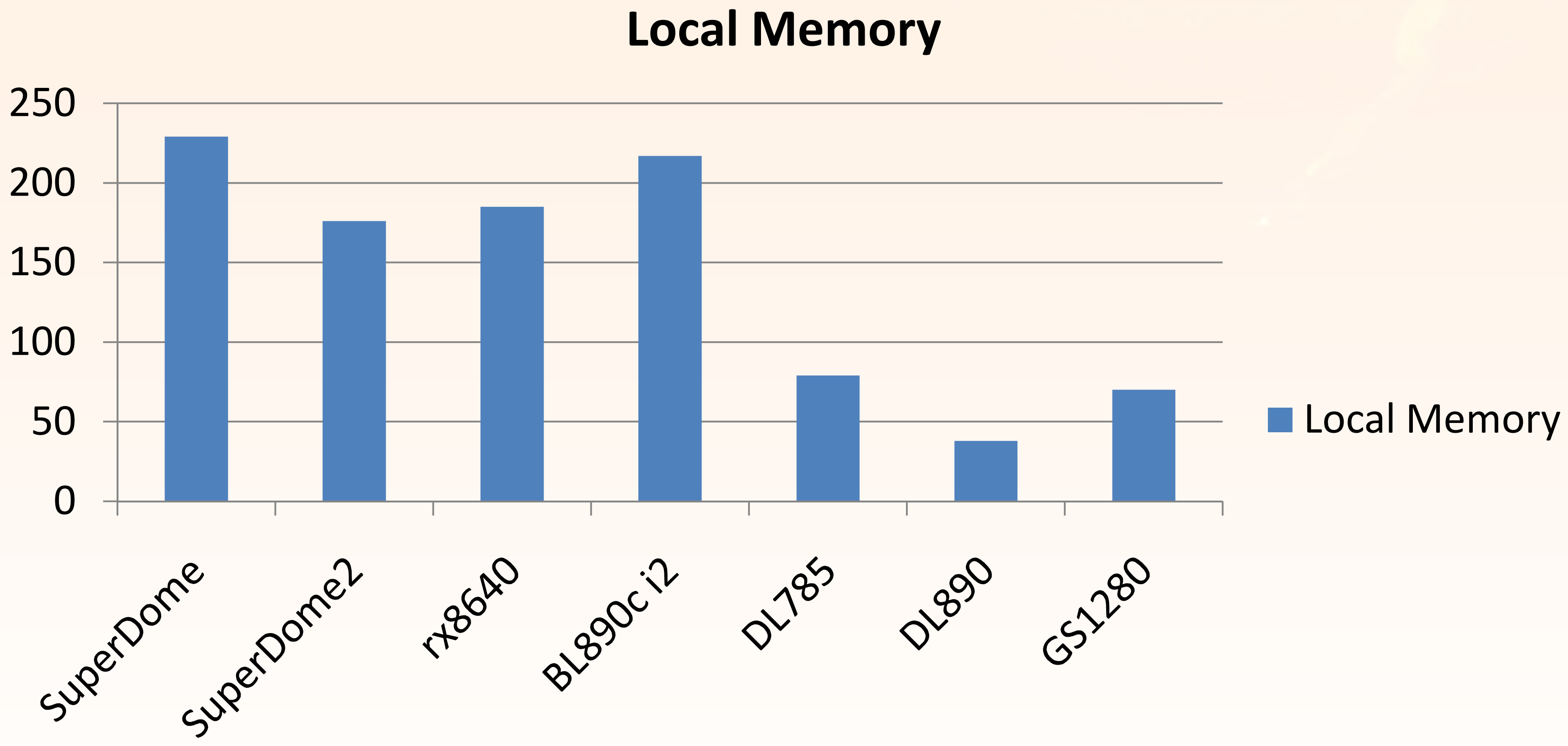
processor7 Q2 (pairs C&D)
processor7 Q3 (pairs E&F)
processor7 Q1 (pairs A&B)

Second Memory Interleaving Domain

Superdome2 Overview



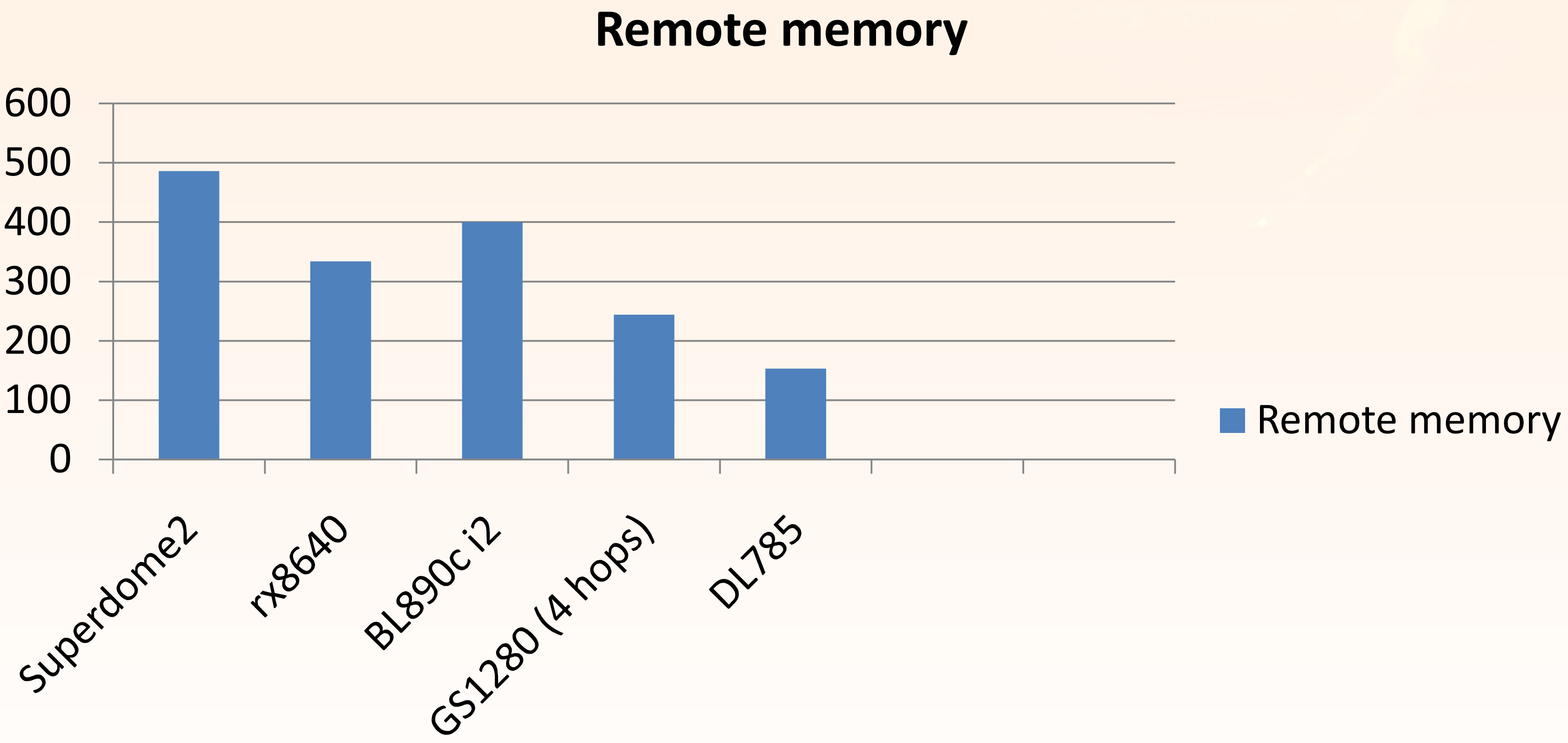
Local Memory Latency



Memory Latency (ns) – Less is better

MAKLEE

Remote Memory Latency



Memory Latency (ns) – Less is better

Latency on the BL890c i2

- Memory latency
  - Inside interleaving domain

Local latency	217 nsec
Latency to a 2nd processor in same blade	288 nsec
Latency to a processor in 2nd blade	300 nsec

  - Across interleaving domains

Latency to direct path processor	300 nsec
Latency to processor in other blade	400 nsec

- Memory latency is not as good as we used to on the Alpha.
- Applications should be tuned to utilize local memory as much as possible.

Local Vs. Interleaved Memory

- Challenges of NUMA based servers:
  - Some CPUs may have an advantage acquiring spinlocks.
  - Some CPUs may have an advantage acquiring locks.
  - Inconsistent performance
    - Performance may change based on the CPU a process is scheduled to.

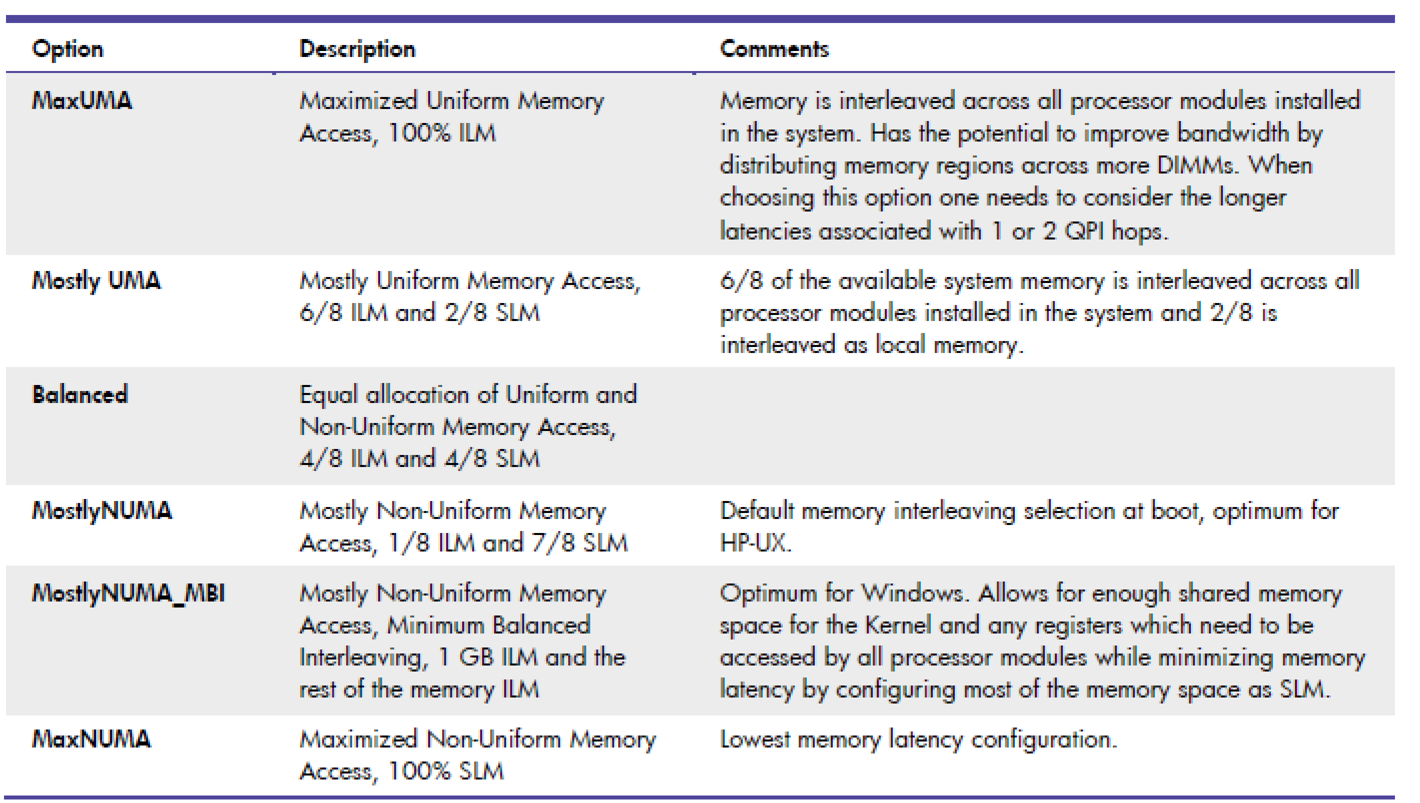
- What could be done to make life a little more fair?
  - Make sure an application is running close to its memory.
    - For example, the dedicated lock manager needs to run close to the lock manager spinlock.
    - Oracle server processes need to run close to the SGA.
  - When the memory footprint of the application is high (shared memory sections than span over more than one domain), consider using Interleaved memory.

  - Until VMS V8.4, VMS only supported interleaved memory.
    - OpenVMS became NUMA aware again (Integrity) starting with OpenVMS V8.4

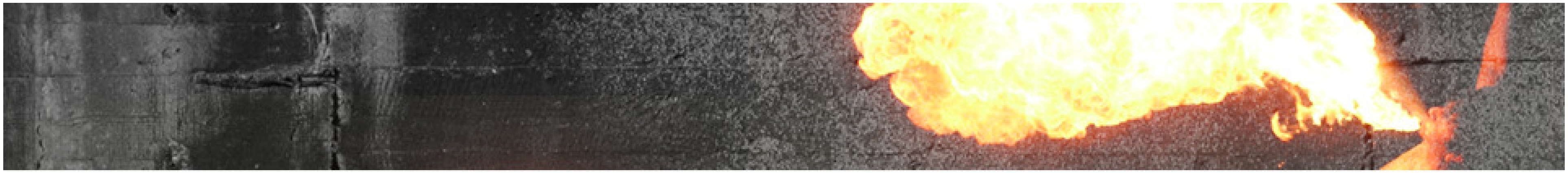
MAKLEE

MEMCONFIG

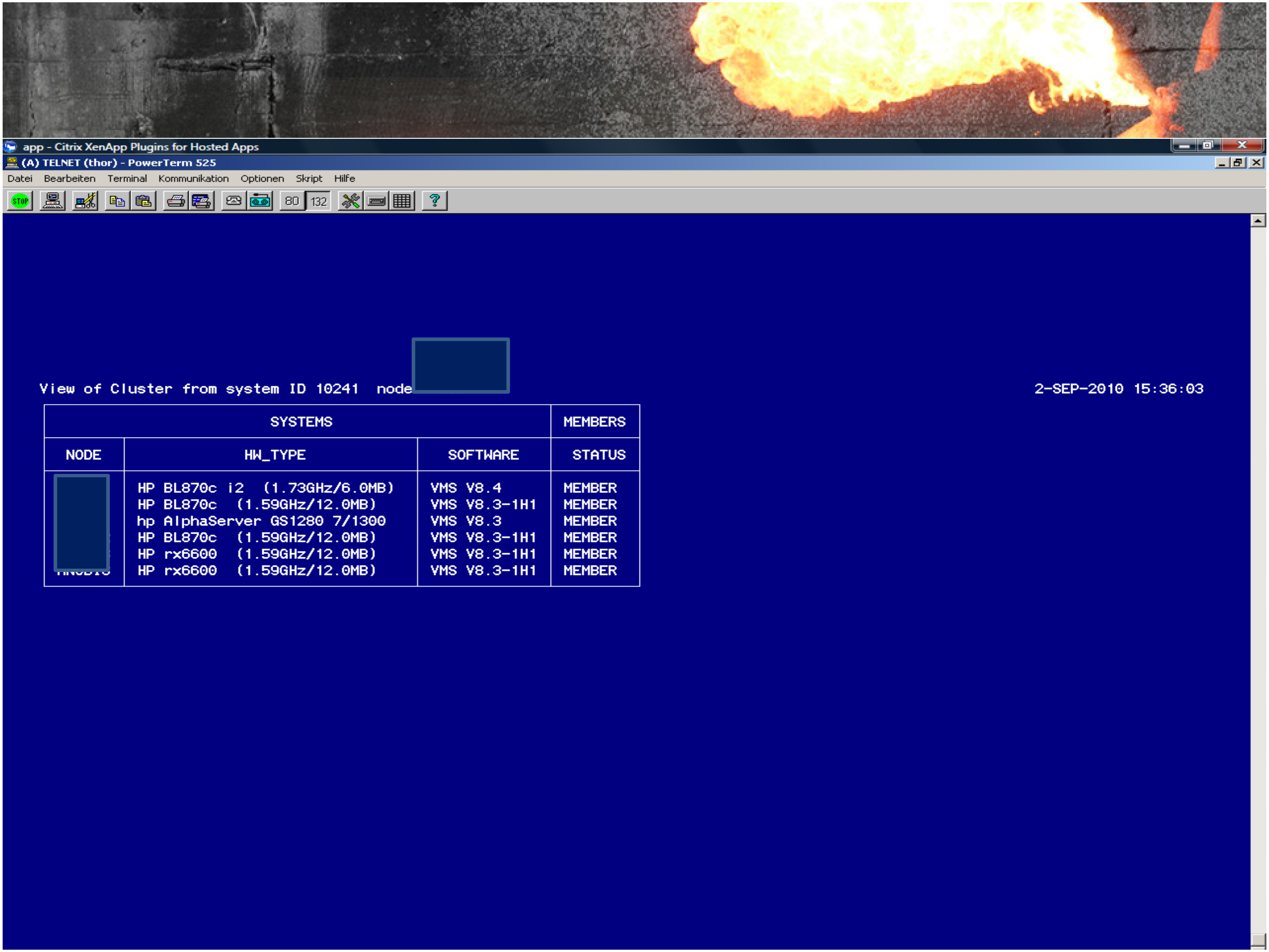
- When migrating to the new BL890c i2, need to decide on memory management policy. Use the EFI MEMCONFIG utility.



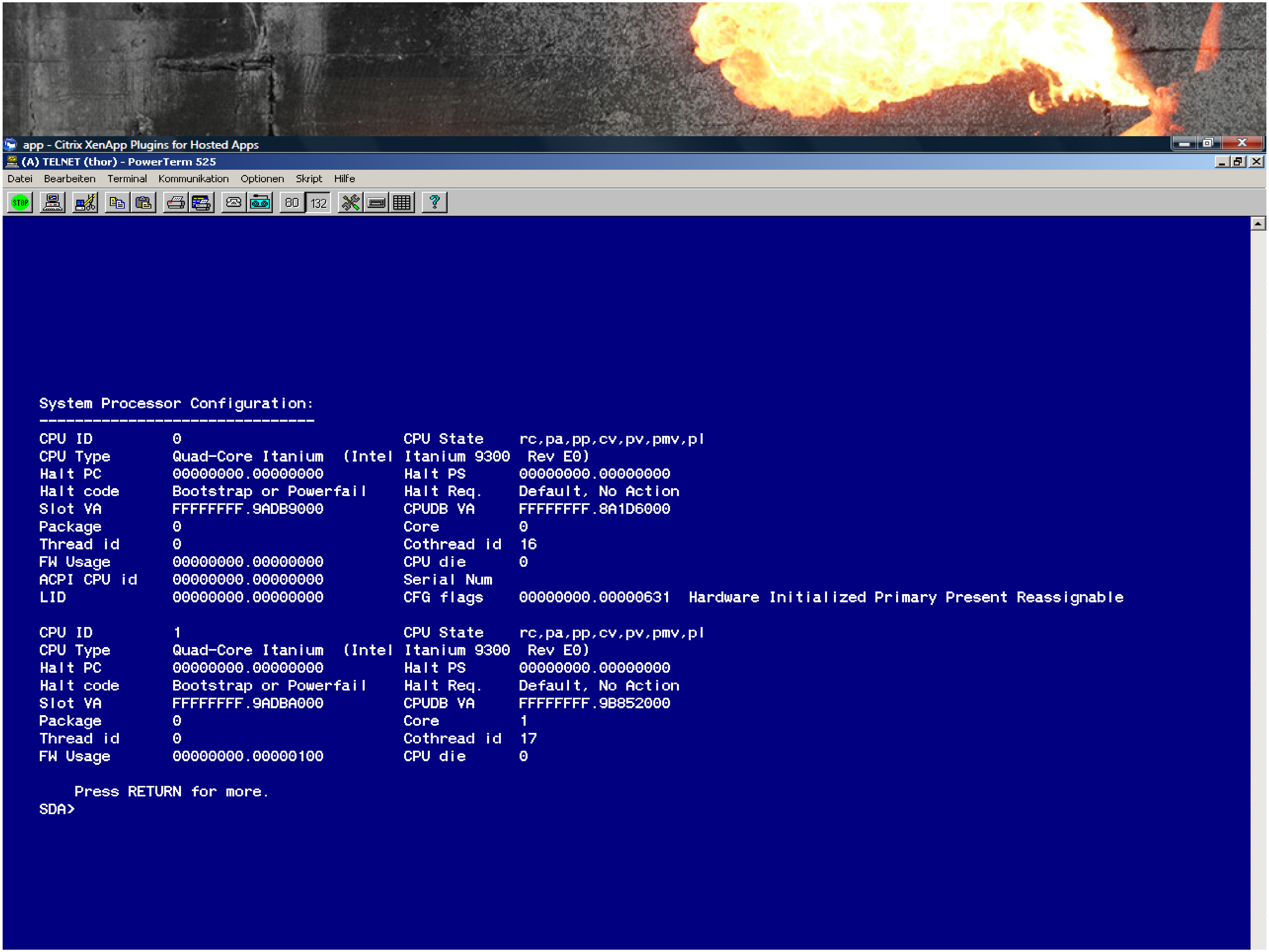
Option	Description	Comments
MaxUMA	Maximized Uniform Memory Access, 100% ILM	Memory is interleaved across all processor modules installed in the system. Has the potential to improve bandwidth by distributing memory regions across more DIMMs. When choosing this option one needs to consider the longer latencies associated with 1 or 2 QPI hops.
Mostly UMA	Mostly Uniform Memory Access, 6/8 ILM and 2/8 SLM	6/8 of the available system memory is interleaved across all processor modules installed in the system and 2/8 is interleaved as local memory.
Balanced	Equal allocation of Uniform and Non-Uniform Memory Access, 4/8 ILM and 4/8 SLM	
MostlyNUMA	Mostly Non-Uniform Memory Access, 1/8 ILM and 7/8 SLM	Default memory interleaving selection at boot, optimum for HP-UX.
MostlyNUMA_MBI	Mostly Non-Uniform Memory Access, Minimum Balanced Interleaving, 1 GB ILM and the rest of the memory ILM	Optimum for Windows. Allows for enough shared memory space for the Kernel and any registers which need to be accessed by all processor modules while minimizing memory latency by configuring most of the memory space as SLM.
MaxNUMA	Maximized Non-Uniform Memory Access, 100% SLM	Lowest memory latency configuration.



OpenVMS Implementation


app - Citrix XenApp Plugins for Hosted Apps
(A) TELNET (thor) - PowerTerm 525
Datei Bearbeiten Terminal Kommunikation Optionen Skript Hilfe
STOP
80
132
?

```
View of Cluster from system ID 10241  node [redacted]                                     2-SEP-2010 15:36:03
+-----------------------------------------------------------+---------+
|                          SYSTEMS                          | MEMBERS |
+---------+----------------------------------+--------------+---------+
|  NODE   |             HW_TYPE              |   SOFTWARE   | STATUS  |
+---------+----------------------------------+--------------+---------+
|[redacted]| HP BL870c i2  (1.73GHz/6.0MB)    | VMS V8.4     | MEMBER  |
|[redacted]| HP BL870c  (1.59GHz/12.0MB)      | VMS V8.3-1H1 | MEMBER  |
|[redacted]| hp AlphaServer GS1280 7/1300     | VMS V8.3     | MEMBER  |
|[redacted]| HP BL870c  (1.59GHz/12.0MB)      | VMS V8.3-1H1 | MEMBER  |
|[redacted]| HP rx6600  (1.59GHz/12.0MB)      | VMS V8.3-1H1 | MEMBER  |
|[redacted]| HP rx6600  (1.59GHz/12.0MB)      | VMS V8.3-1H1 | MEMBER  |
+---------+----------------------------------+--------------+---------+
```


app - Citrix XenApp Plugins for Hosted Apps
(A) TELNET (thor) - PowerTerm 525
Datei
Bearbeiten
Terminal
Kommunikation
Optionen
Skript
Hilfe
STOP
80
132
?

```
System Processor Configuration:
-------------------------------
CPU ID         0                         CPU State    rc,pa,pp,cv,pv,pmv,pl
CPU Type       Quad-Core Itanium  (Intel Itanium 9300  Rev E0)
Halt PC        00000000.00000000         Halt PS      00000000.00000000
Halt code      Bootstrap or Powerfail    Halt Req.    Default, No Action
Slot VA        FFFFFFFF.9ADB9000         CPUDB VA     FFFFFFFF.8A1D6000
Package        0                         Core         0
Thread id      0                         Cothread id  16
FW Usage       00000000.00000000         CPU die      0
ACPI CPU id    00000000.00000000         Serial Num
LID            00000000.00000000         CFG flags    00000000.00000631  Hardware Initialized Primary Present Reassignable

CPU ID         1                         CPU State    rc,pa,pp,cv,pv,pmv,pl
CPU Type       Quad-Core Itanium  (Intel Itanium 9300  Rev E0)
Halt PC        00000000.00000000         Halt PS      00000000.00000000
Halt code      Bootstrap or Powerfail    Halt Req.    Default, No Action
Slot VA        FFFFFFFF.9ADBA000         CPUDB VA     FFFFFFFF.9B852000
Package        0                         Core         1
Thread id      0                         Cothread id  17
FW Usage       00000000.00000100         CPU die      0

    Press RETURN for more.
SDA>
```

RAD_SUPPORT

RAD_SUPPORT

(Alpha only) RAD_SUPPORT enables RAD-aware code to be executed on systems that support Resource Affinity Domains (RADs); for example, AlphaServer GS160 systems. A RAD is a set of hardware components (CPUs, memory, and I/O) with common access characteristics.

Bits are defined in the RAD_SUPPORT parameter as follows:

RAD_SUPPORT (default is 79; bits 0-3 and 6 are set)

```
 3   2 2   2 2         1 1
 1   8 7   4 3         6 5         8 7         0
+-----+-----+-----------+-----------+-----------+
|00|00| skip|ss|gg|ww|pp|00|00|00|00|0p|df|cr|ae|
+-----+-----+-----------+-----------+-----------+
```

```
Bit 0 (e): Enable    - Enables RAD support

Bit 1 (a): Affinity  - Enables Soft RAD Affinity (SRA) scheduling
                       Also enables the interpretation of the skip
                       bits, 24-27.

Bit 2 (r): Replicate - Enables system-space code replication
```

RAD_SUPPORT

```
Bit 3 (c): Copy      - Enables copy on soft fault

Bit 4 (f): Fault     - Enables special page fault allocation
                       Also enables the interpretation of the
                       allocation bits, 16-23.

Bit 5 (d): Debug     - Reserved to HP

Bit 6 (p): Pool      - Enables per-RAD non-paged pool

Bits 7-15:           - Reserved to HP

Bits 16-23:          - If bit 4 is set, bits 16-23 are interpreted
                       as follows:

Bits 16,17 (pp): Process = Pagefault on process (non global)
                           pages
Bits 18,19 (ww): Swapper = Swapper's allocation of pages for
                           processes
Bits 20,21 (gg): Global  = Pagefault on global pages
Bits 22,23 (ss): System  = Pagefault on system space pages
```

VMS representation MostlyNUMA

```
$@sys$examples:rad

Node: XXXX Version: V8.4      System: HP BL870c i2  (1.73GHz/6.0MB)

RAD   Memory (GB)   CPUs
===   ===========   ===============
  0       28.00     0-3,16-19
  1       28.00     8-11,24-27
  2       28.00     4-7,20-23
  3       28.00     12-15,28-31
  4       15.99     0-31
```

SDA SHOW PFN

```
sda> show pfn/rad


Page RAD summary
----------------

         RAD          Free pages          Zeroed pages
         ---         ------------         ------------
        0000                    0                    0
        0001               233783                65535
        0002              3538223                    1 *
        0003              3395833              3396682
        0004                    0                    0

There are -3247242 additional pages in the free list

* An error occurred scanning this list
  The count of additional pages given may not be correct
SDA>
```

show rad/pxml

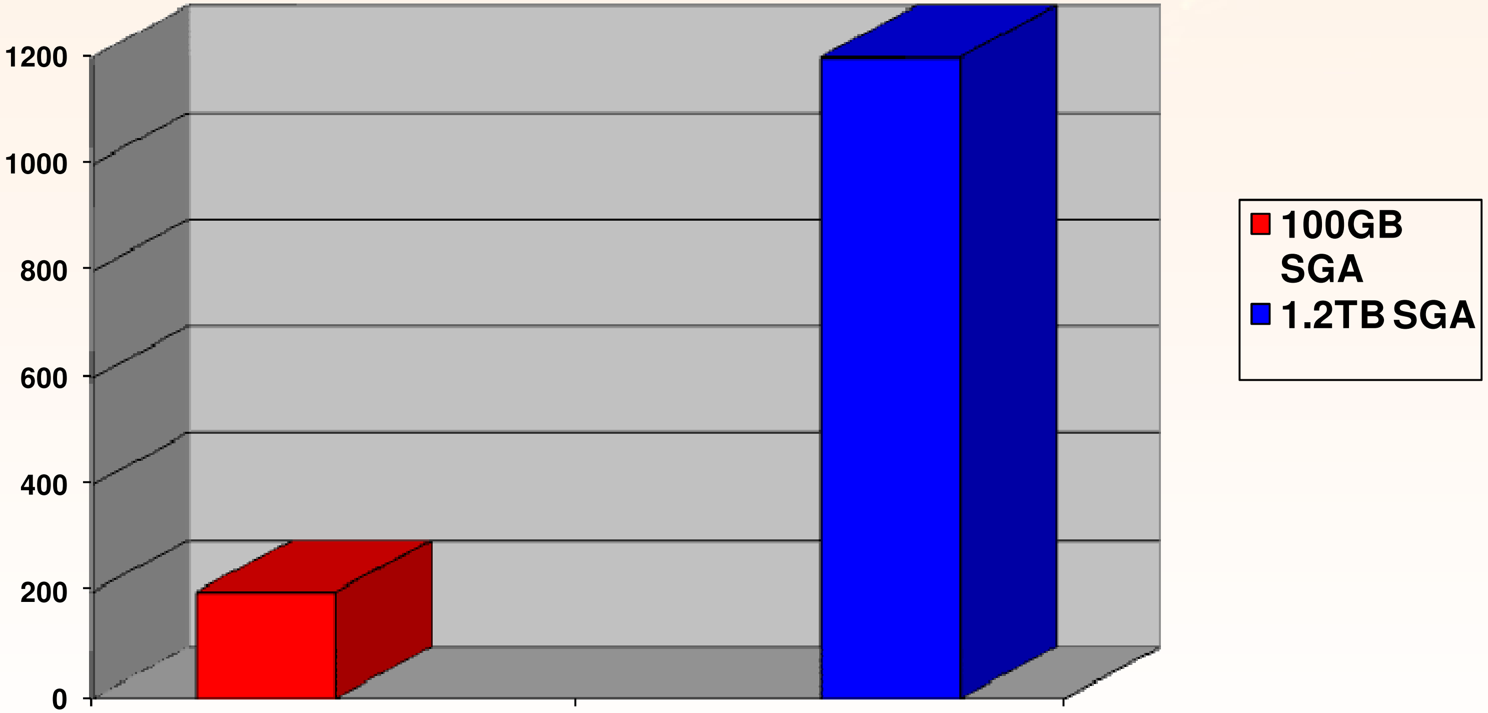
```
Locality #03 (RAD #02)
----------------------

  Address:        FFFFF802.ECF22788   Size:                       000000D8
  Average:                 000001B0   Spread:                     000016AA
  Base RAD:                      04

  CPU count:               00000008   CPU bitmap:        00000000.00F000F0

  Memory range(s):         00000001   00000020.00000000:00000026.FFFFFFFF

                         (as PFNs)    00000000.01000000:00000000.0137FFFF

  Total memory:   00000007.00000000   (28672MB)

  RAD preference array:               00000002 00000004 00000003 00000000
                                      00000001
```

VMS

- Use the SHOW FASTPATH command and move device interrupts to the low numbered CPUs.
- Move the dedicated lock manager close to the lock manager spin lock.
- Move TCP/IP close to the TCP/IP spinlock.
- Memory sections
  - Use the /RAD qualifier allocating reserved memory from a specific RAD.
    - mc sysman add reserved_section_name /rad=x
  - Use interleaved memory for shared memory sections that span over one RAD.
    - Makes sense also for systems running a single Oracle database
    - /rad=4 pn BL890c i2
  - Use local memory for small memory sections.
- Experiment with RAD_SUPPORT
  - No documentation as to what is happening under the hood, disable if interleaved memory is used, reduce unnecessary overhead in MMG

Now…Can you explain it??

- Superdome2 , 2TB RAM, HP-UX 11.31 (update 7)
- Oracle 11gR2



Elapsed time (ms) to execute single select query