Почему IBM POWER8 оптимальная платформа для PostgreSQL

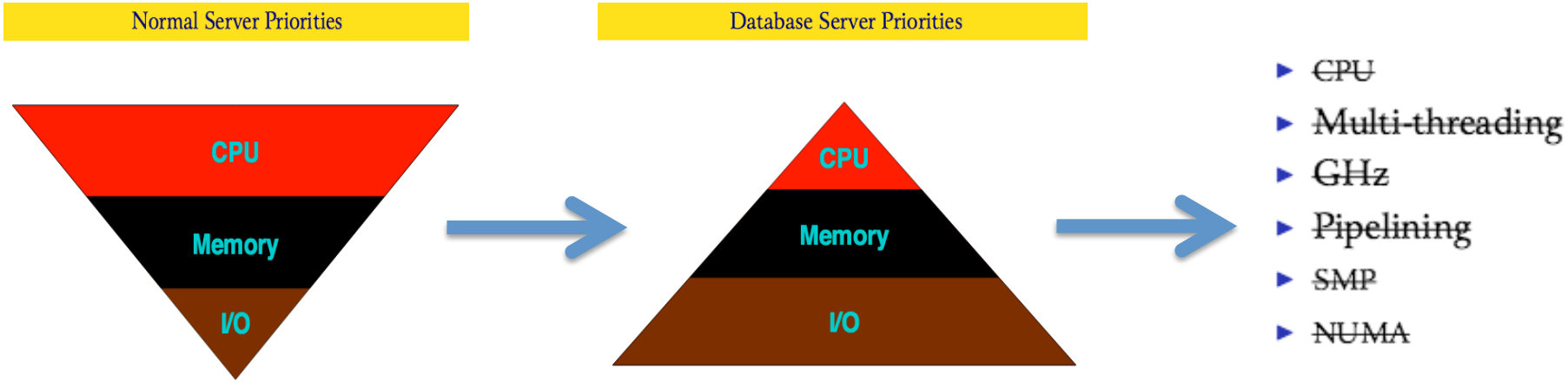
Иван Гончаров

Технический специалист

igoncharov@ru.ibm.com

What server should I choose for PG?

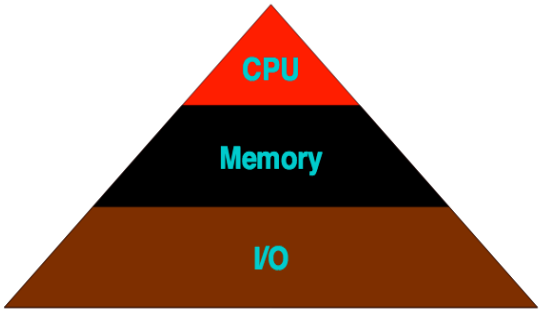
Old-fashioned approach



Slides borrowed from Bruce Momjian: http://momjian.us/main/writings/pgsql/hw_selection.pdf

Why it is still extremely relevant

Database Server Priorities

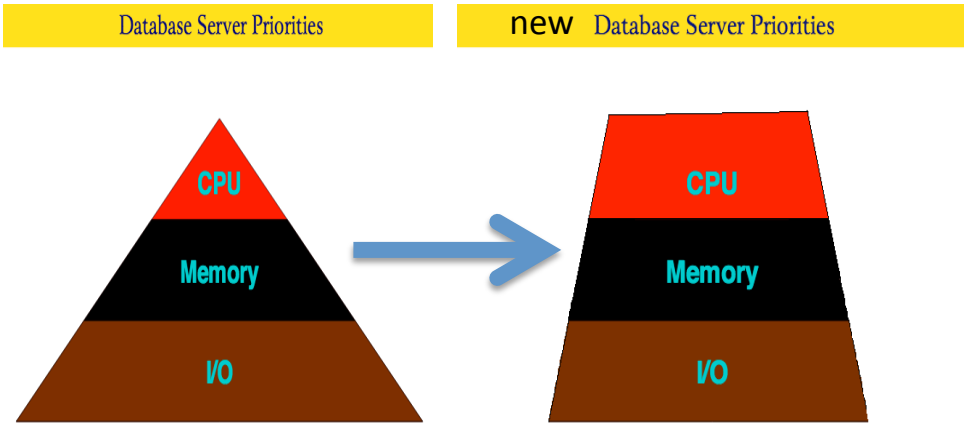


.5	ns	L1 cache reference
5	ns	Branch mispredict
7	ns	L2 cache reference
25	ns	Mutex lock/unlock
100	ns	Main memory reference
250,000	ns	Read 1 MB sequentially from memory
10,000,000	ns	Disk seek
20,000,000	ns	Read 1 MB sequentially from disk

Five orders of magnitude difference

Slides from: http://momjian.us/main/writings/pgsql/hw_selection.pdf
https://gist.github.com/jboner/2841832

Industry trends

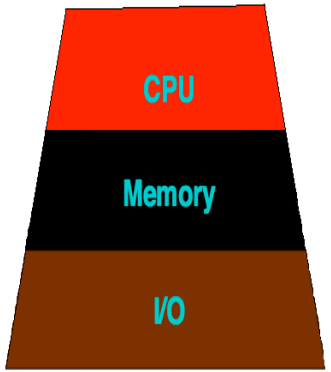


- Virtualization
  - Many VMs per CPU
- More and more RAM every year
  - 32TB per server available
- Flash memory everywhere
  - 1M IO/s with 0.1ms latency is not Sci-Fi

Slides from: http://momjian.us/main/writings/pgsql/hw_selection.pdf

2010 to 2016 difference

new Database Server Priorities



.5	ns	L1 cache reference
5	ns	Branch mispredict
7	ns	L2 cache reference
25	ns	Mutex lock/unlock
100	ns	Main memory reference
250,000	ns	Read 1 MB sequentially from memory
16,000	ns	small SSD random read
150,000	ns	4k SSD random read
1,000,000	ns	Read 1 MB sequentially from SSD

One order of magnitude difference

Slides from: http://momjian.us/main/writings/pgsql/hw_selection.pdf
https://gist.github.com/jboner/2841832
http://www.eecs.berkeley.edu/~rcs/research/interactive_latency.html

Intel vs. AMD, right?

What is IBM POWER?

17 ноября 2015 в 02:19

PostgreSQL на многоядерных серверах Power 8

PostgreSQL*, Блог компании Postgres Professional

обслуживание мэйнфрейма требует специальных знаний, в отличие от сети дешёвых персоналок.

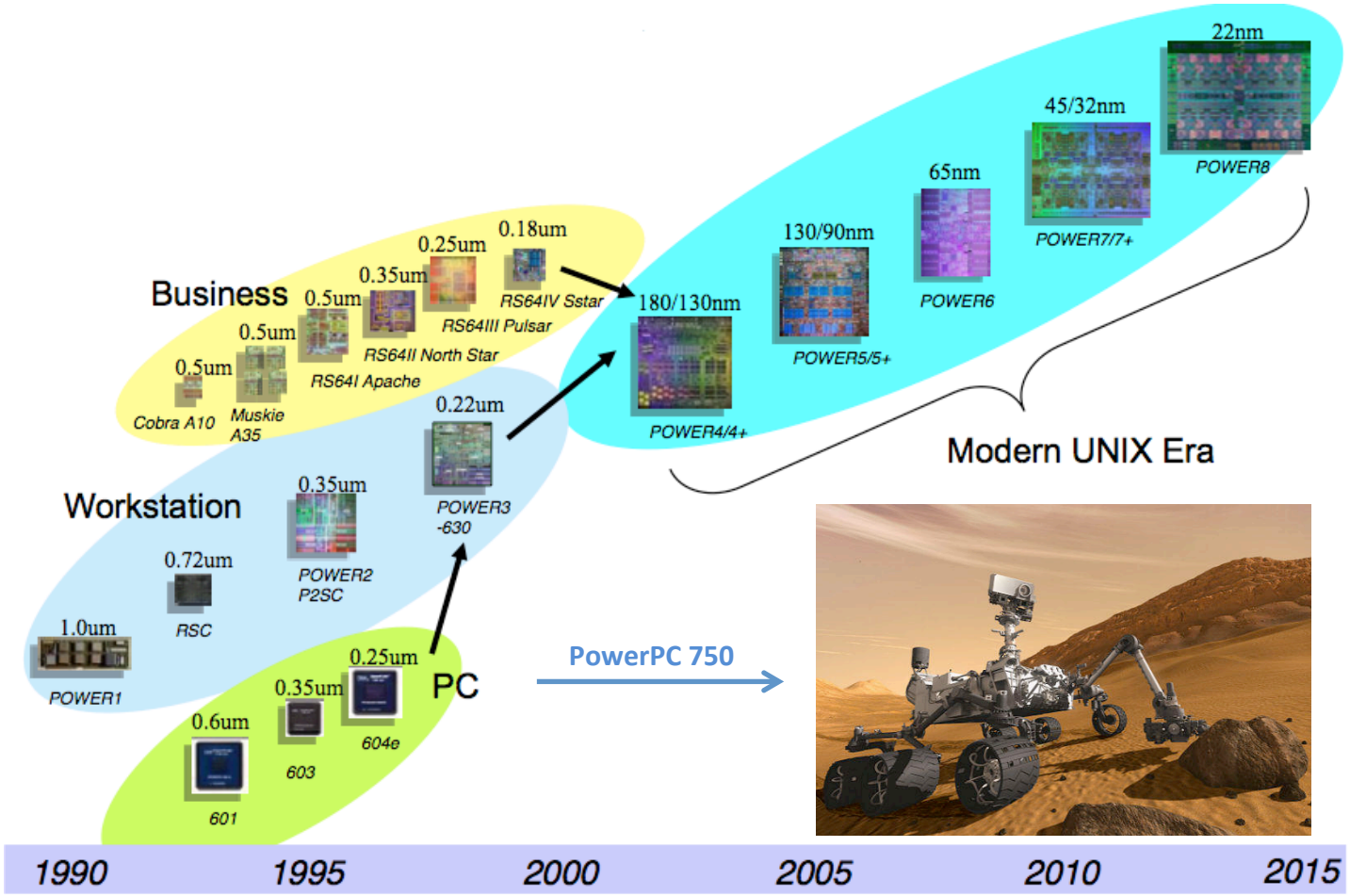
What is IBM POWER?

Mainframe

≠

Power Server

What is IBM POWER?



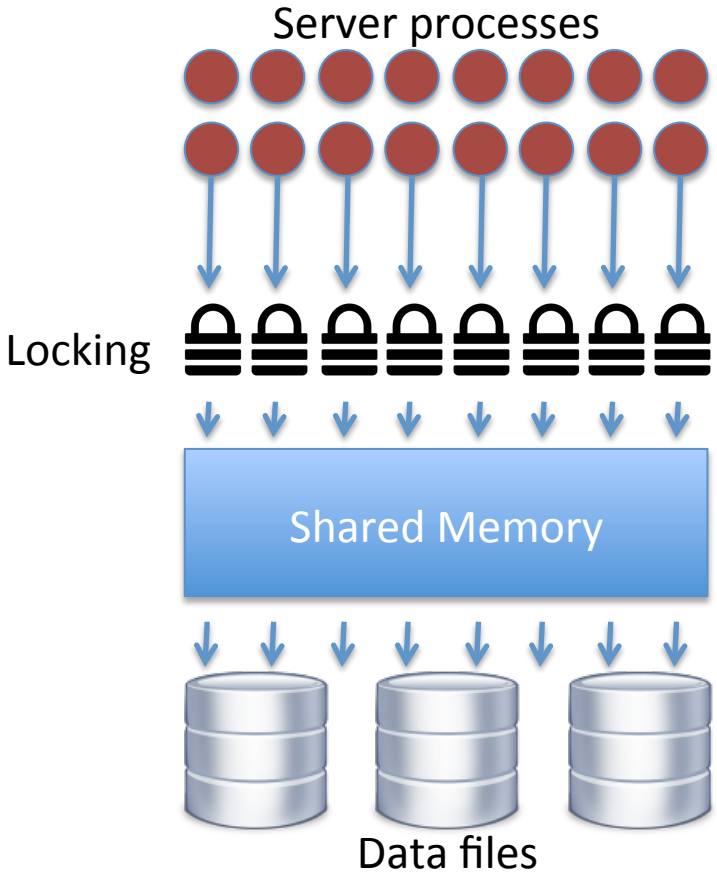
What is IBM POWER?

- 8 -> 192 cores
- 32GB -> 32TB RAM
- Many other differences

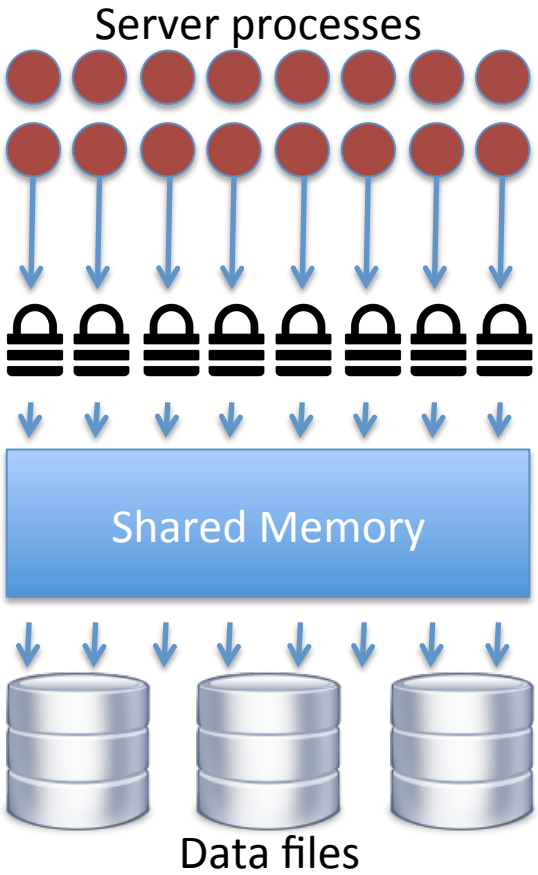
Linux

Why PostgreSQL on POWER?

PostgreSQL DB, server point of view



PostgreSQL DB, server point of view



stress

stress

stress

stress

CPU performance
- Single thread
- Multithread
- Throughput

SMP interconnect
- Bandwidth
- Latency

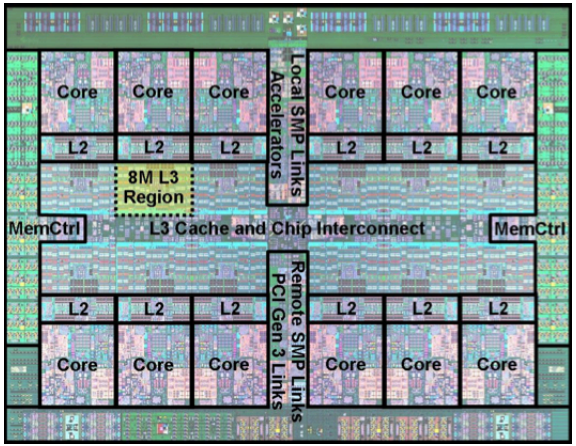
Memory performance
- Bandwidth
- Latency

IO performance
- Bandwidth
- Latency

CPU performance

Single thread performance

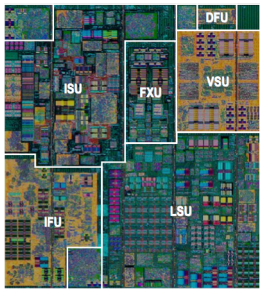
- High operating frequency, 3GHz to 4.5GHz
- Large caches, L1 96KB/core, L2 512KB/core, L3 96MB/chip, L4 128MB/chip (external)



Chip

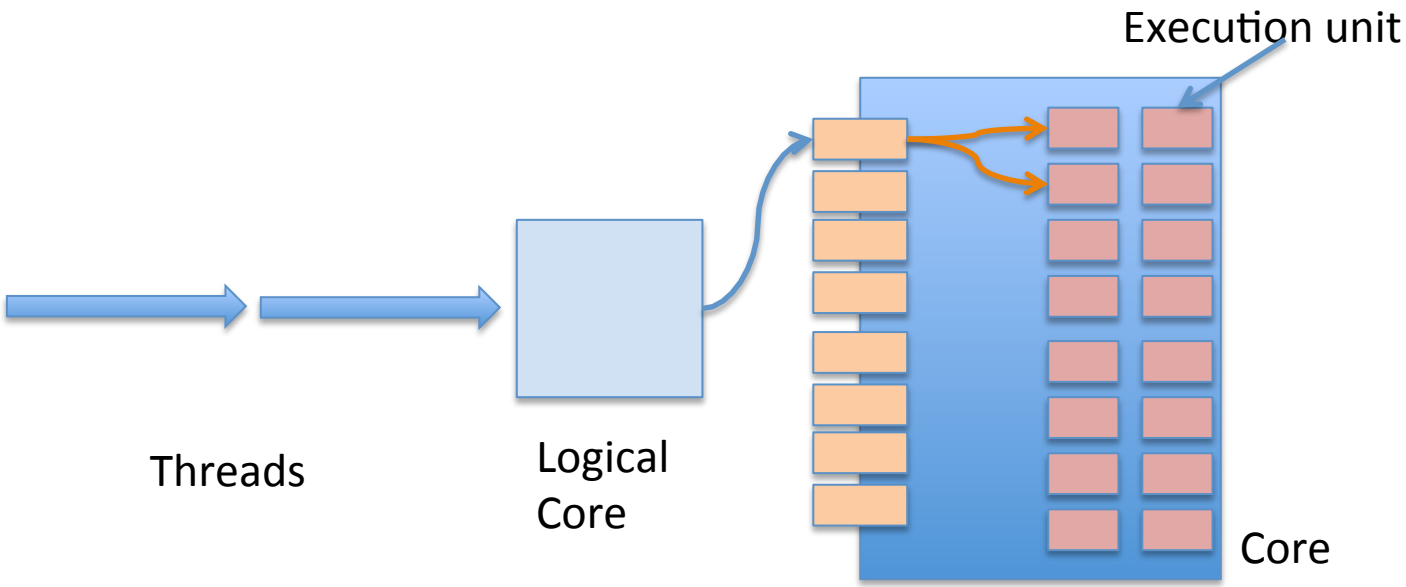
Multithreading

- lots of execution units, 16 exec pipes, 10 issue, 8 dispatch/commit
- Wide cache bandwidth
- large caches

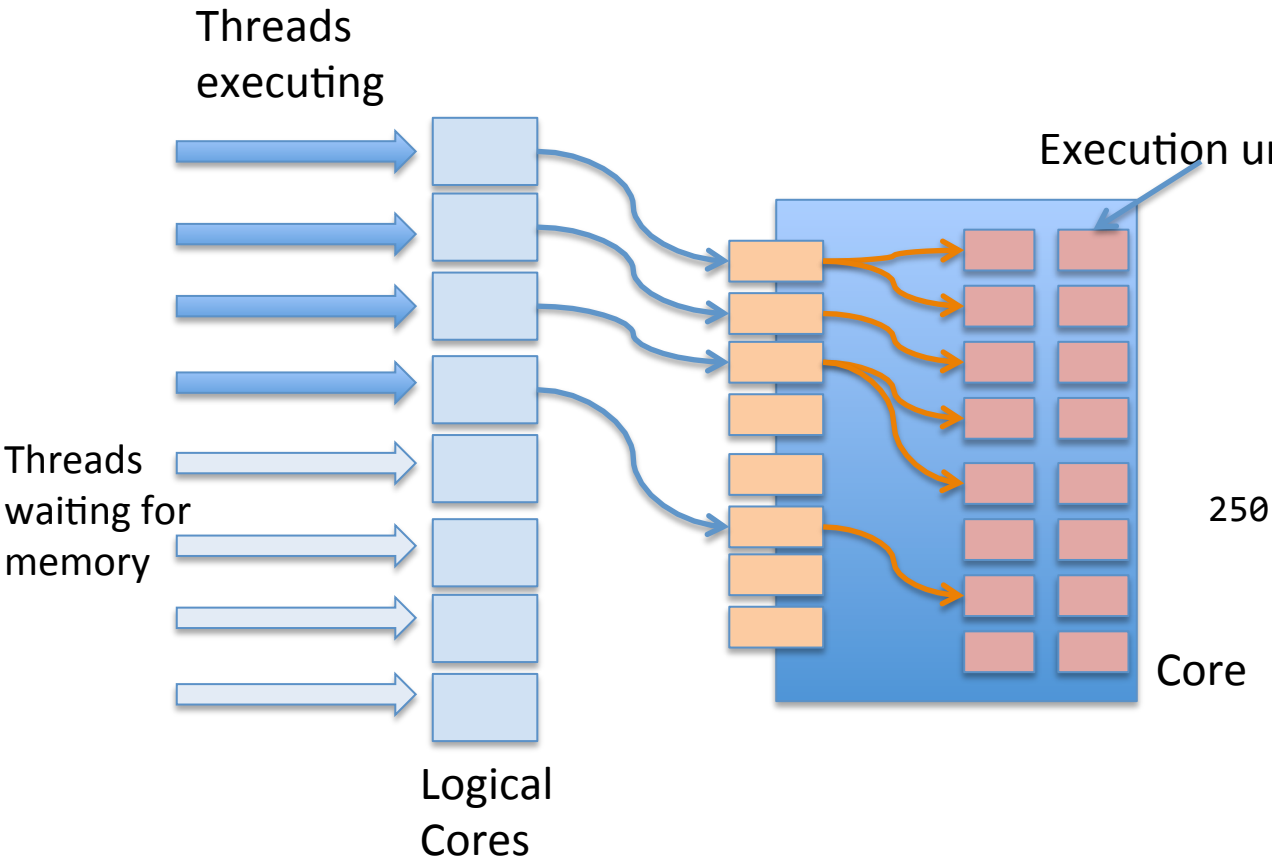


Core

Simultaneous Multi Threading (SMT)

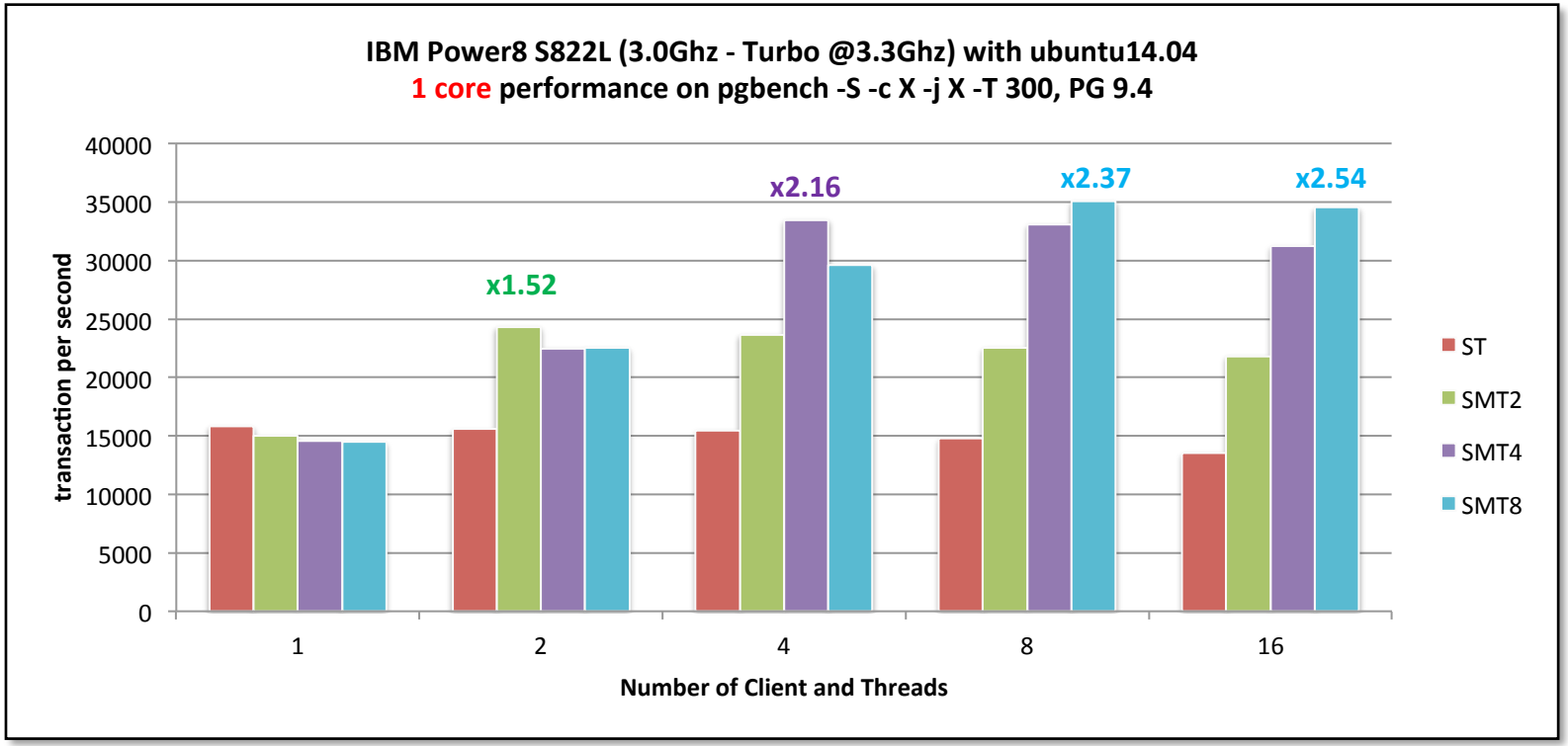


Simultaneous Multi Threading (SMT)

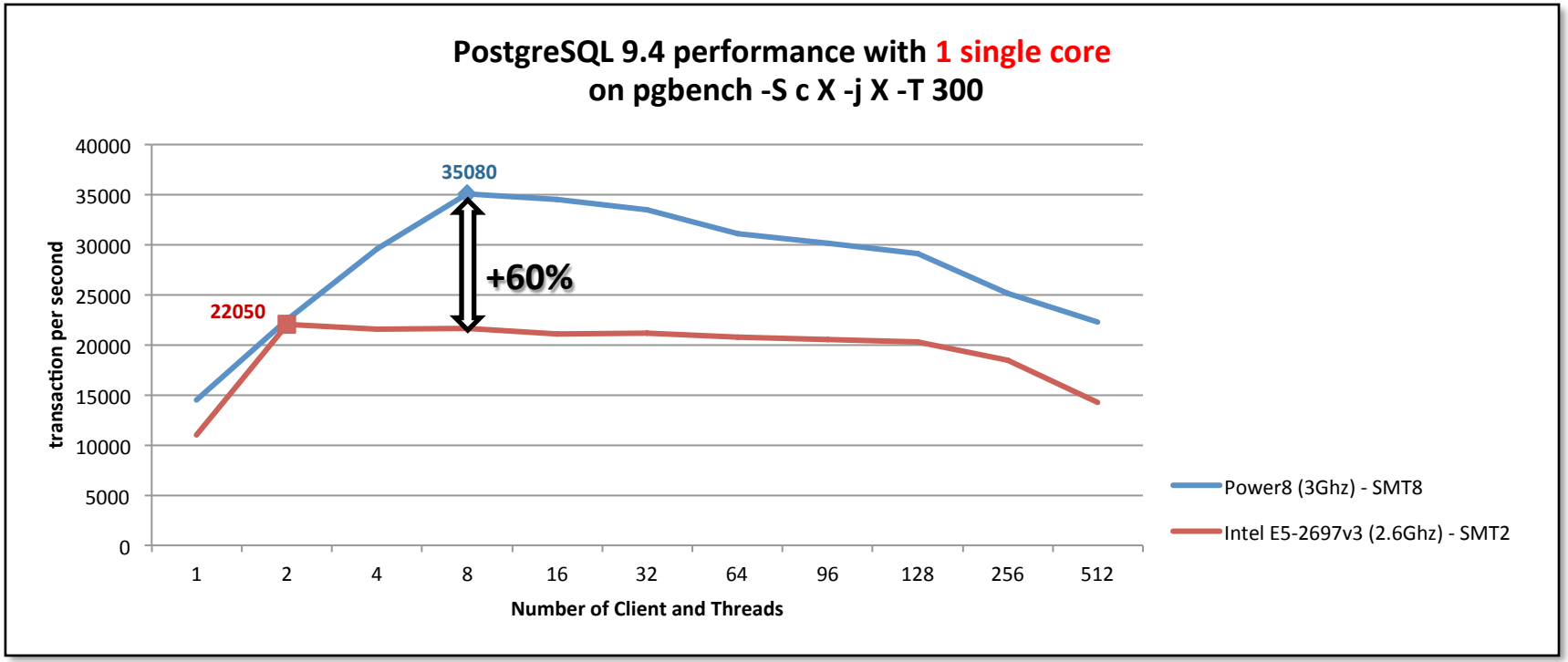



```
    .5  ns L1 cache reference
     5  ns Branch mispredict
     7  ns L2 cache reference
    25  ns Mutex lock/unlock
   100  ns Main memory reference
250,000  ns Read 1 MB from memory
```

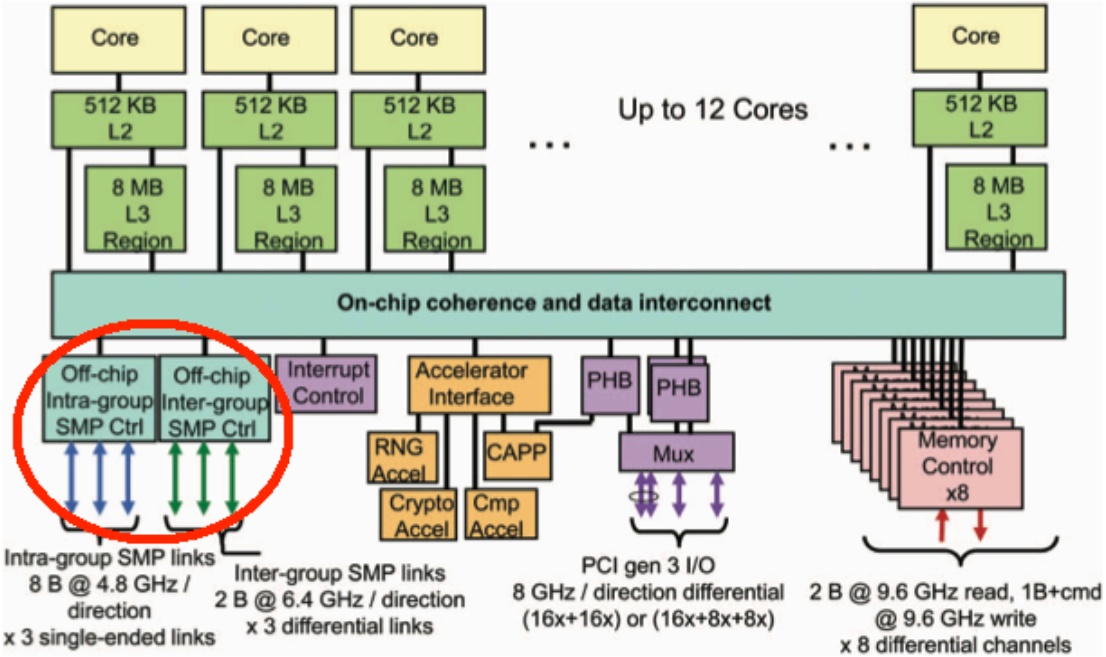
SMT for PostgreSQL, pgbench test



x86 core vs. Power8 core



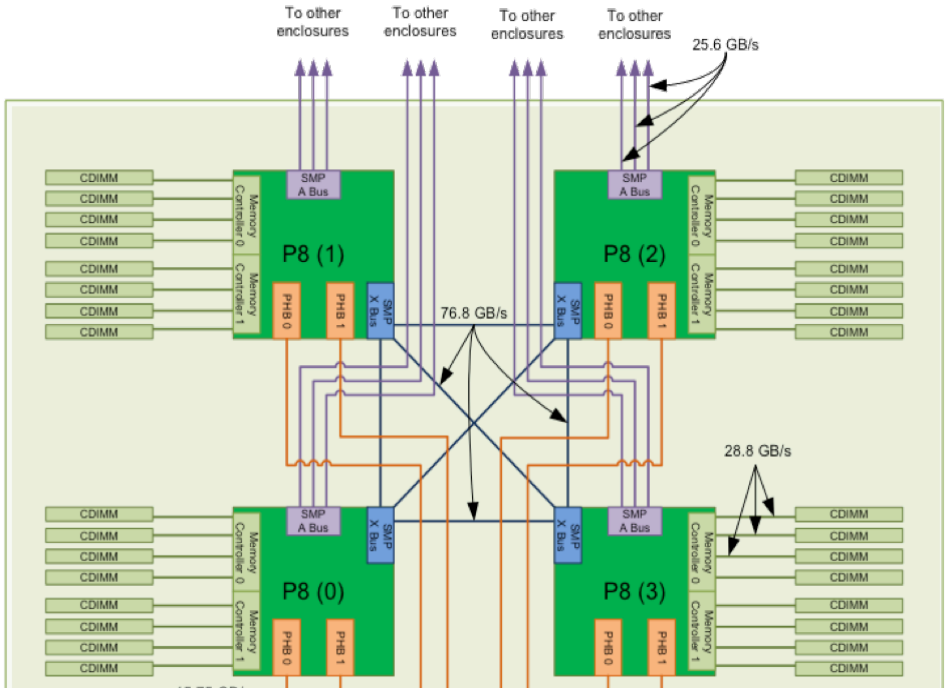
SMP interconnect



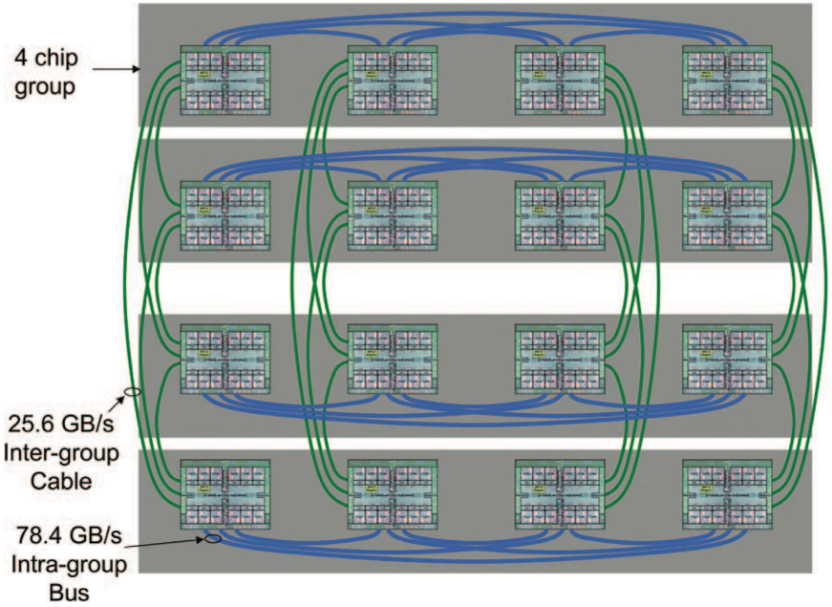
Why it is important

- Locking generating lot of SMP traffic
- Remote memory read/write

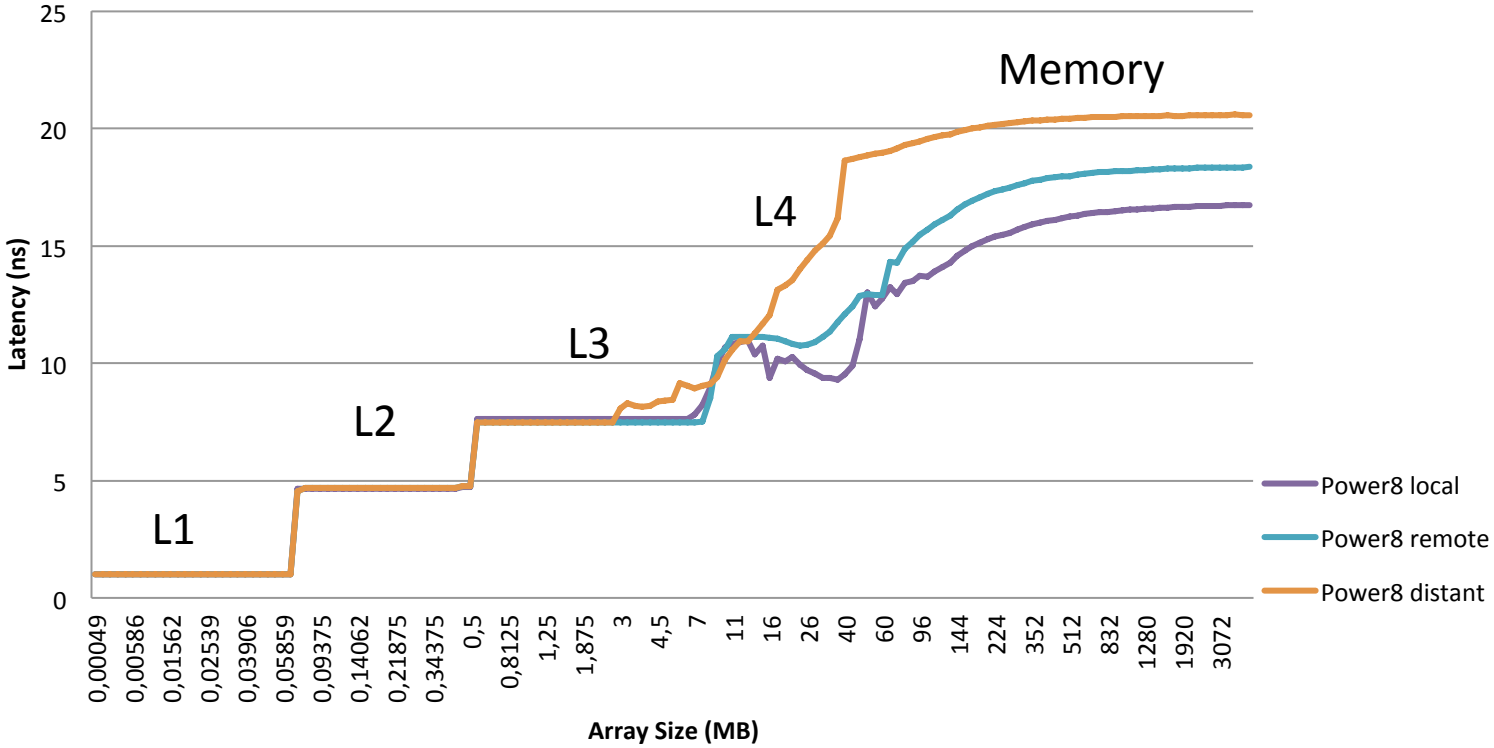
SMP interconnect = scalability



2s	104 GB/s
4s	460 GB/s
16s	2457 GB/s

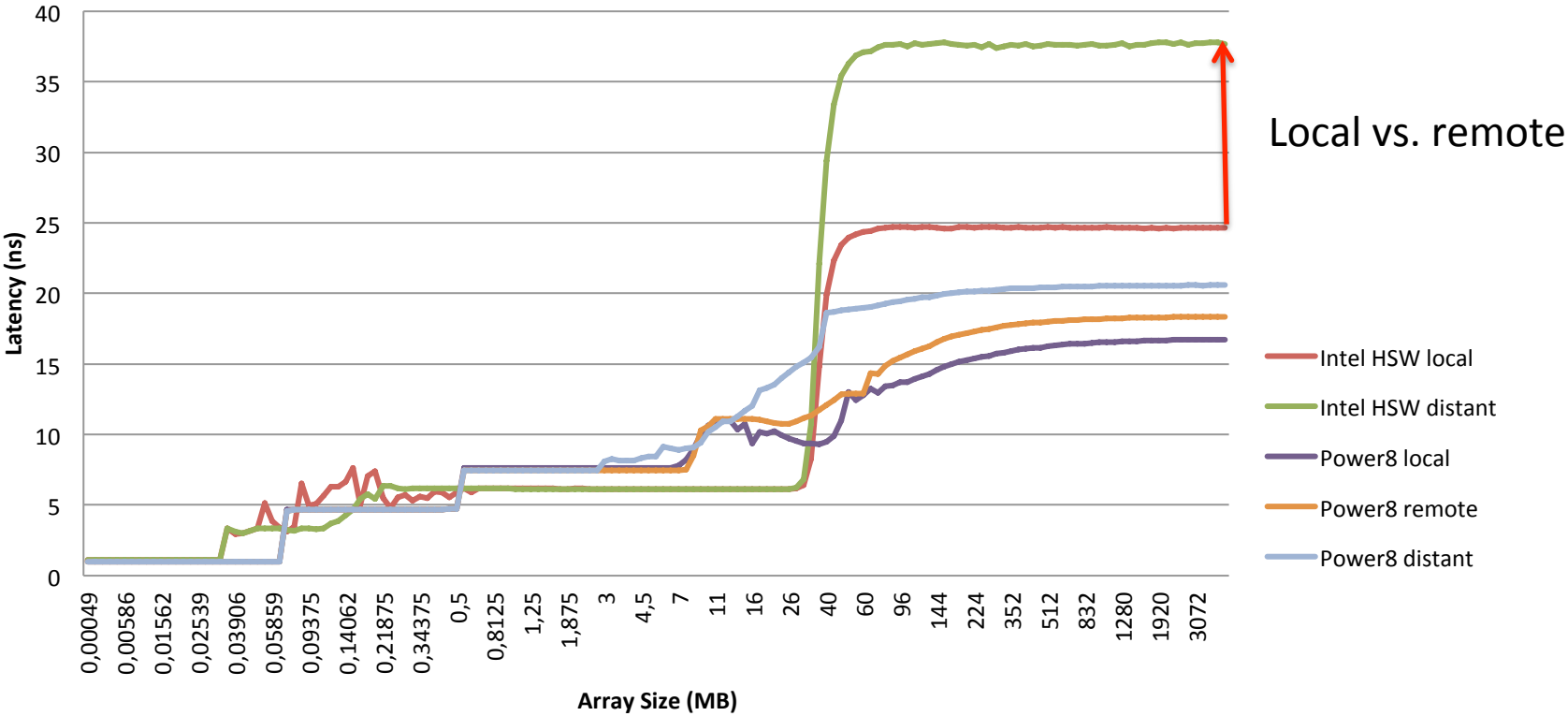


NUMA remote vs. local latency



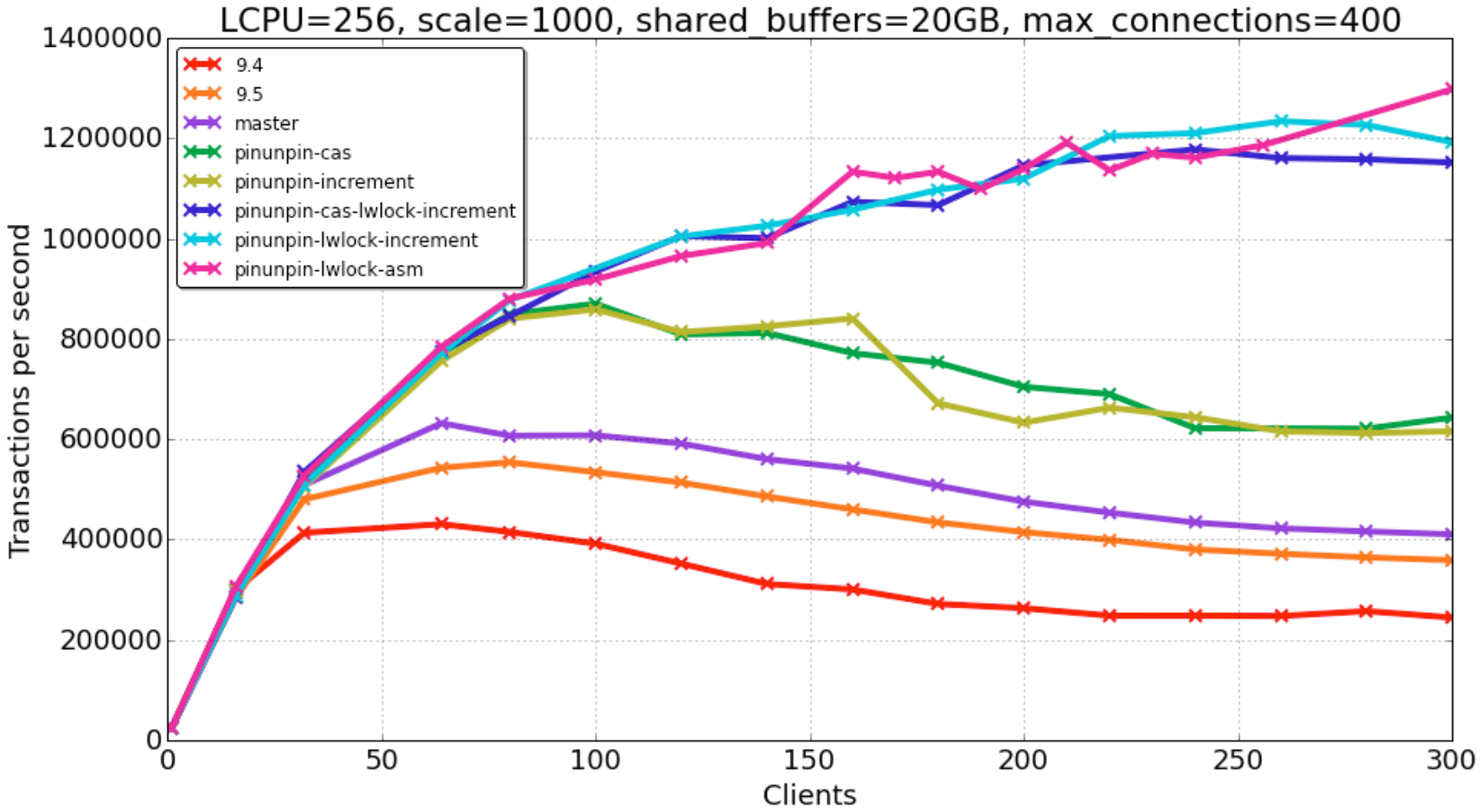
Test performed with lmbench comparing S824 (2 socket - 12 cores) fully populated in memory,
Courtesy of Sebastien Chabrolles, IBM France

NUMA remote vs. local latency



Test performed with lmbench comparing S824 (2 socket - 12 cores) fully populated in memory,
Courtesy of Sebastien Chabrolles, IBM France

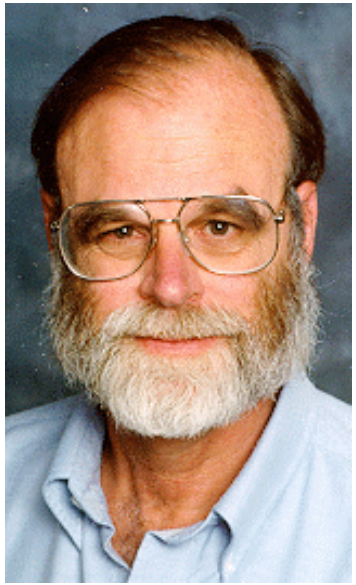
Pgbench scalability test, 32 cores



From: https://habrahabr.ru/company/postgrespro/blog/270827/

Memory

Memory performance

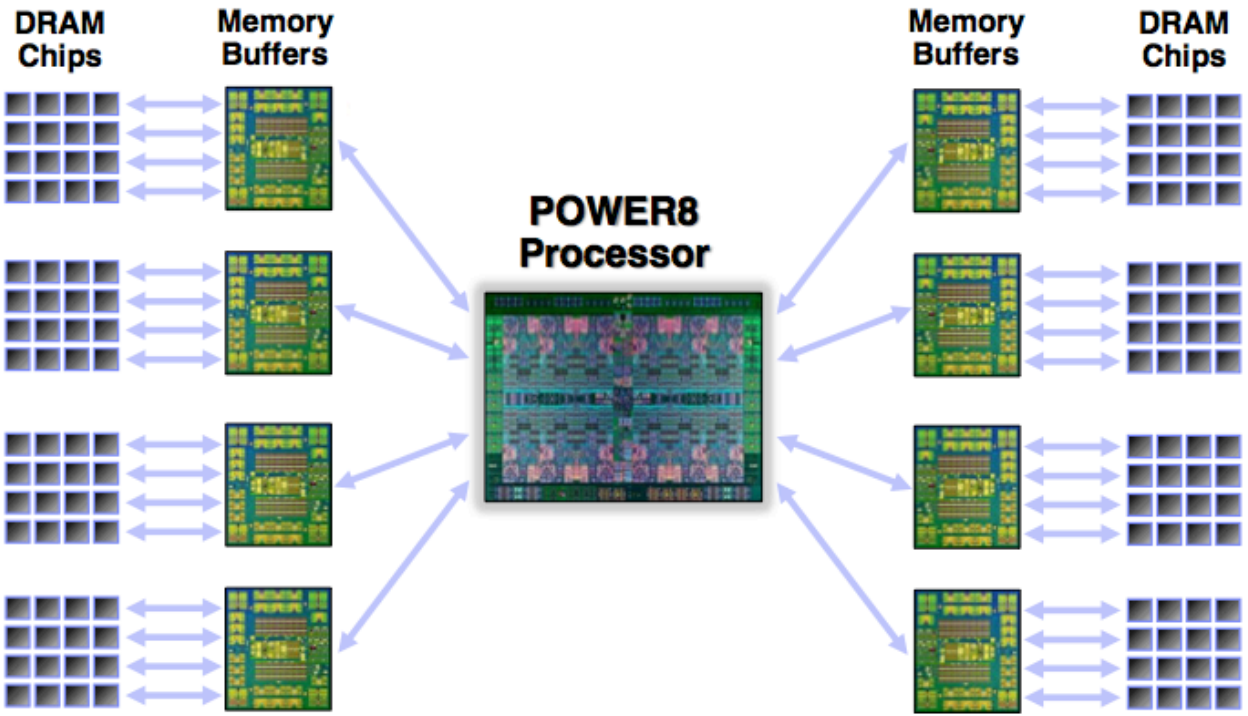


"Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King"

Jim Gray
Microsoft
December 2006

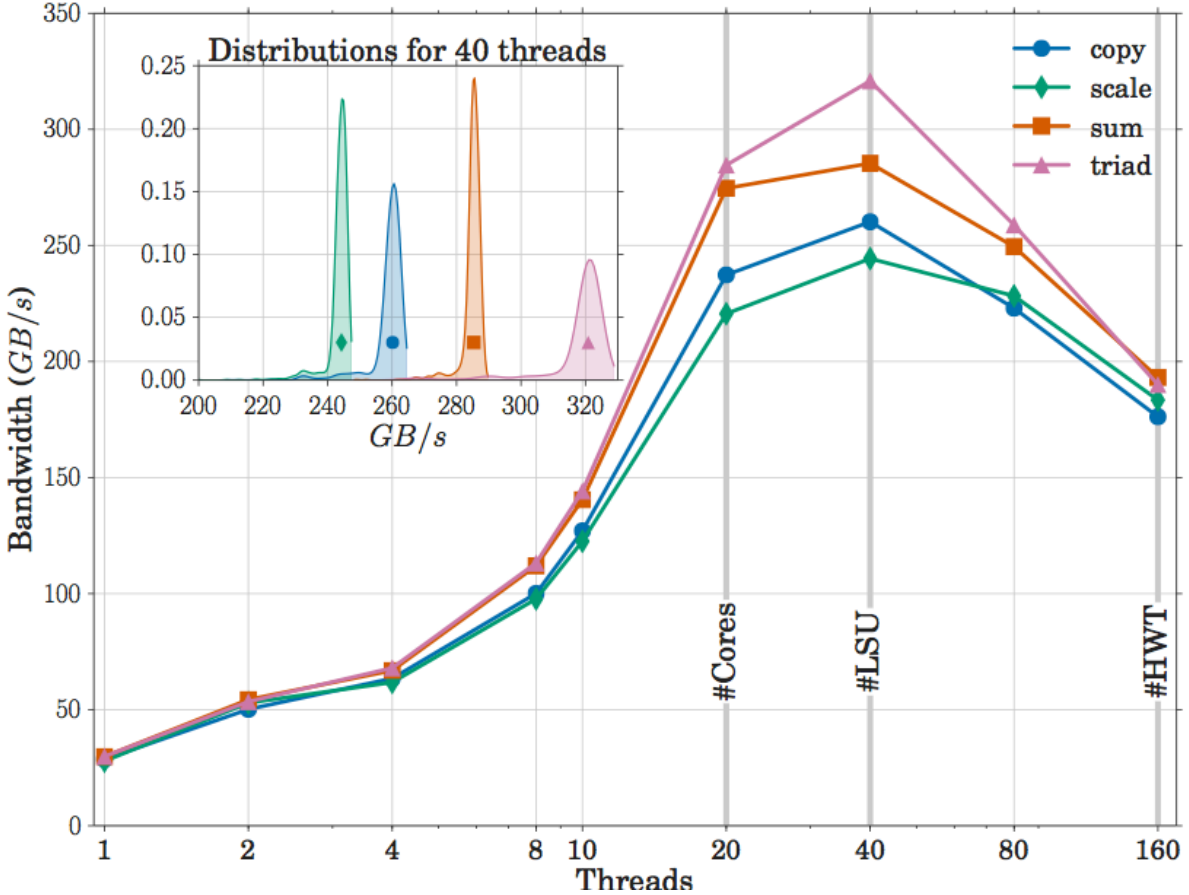
http://research.microsoft.com/en-us/um/people/gray/talks/Flash_is_Good.ppt

POWER8 memory architecture



- Up to 8 high speed channels, each 2B rd + 1B wr at 9.6 GHz for up to 230 GB/s
- Up to 32 total DDR ports yielding 410 GB/s peak at the DRAM
- Up to 128MB L4 cache and 1 TB memory capacity per processor socket

Memory bandwidth, STREAM micro benchmark



S822, 20 cores (2 sockets)

- with 2 streams

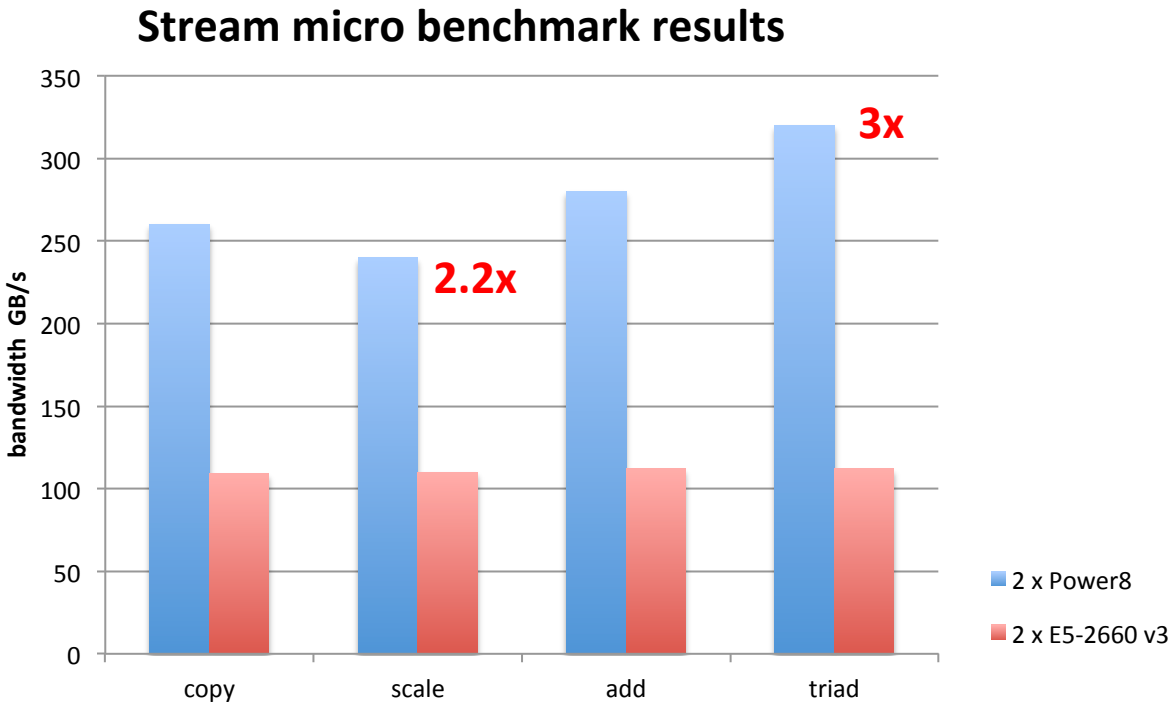
```
copy        c[i] = a[i];
scale       b[i] = s*c[i];
```

- with 3 streams

```
sum         a[i] = b[i]+c[i];
triad       a[i] = s*b[i]+c[i];
```

Source: Performance Evaluation of Scientific Applications on POWER8, by Jülich Supercomputing Centre

Memory bandwidth against x86



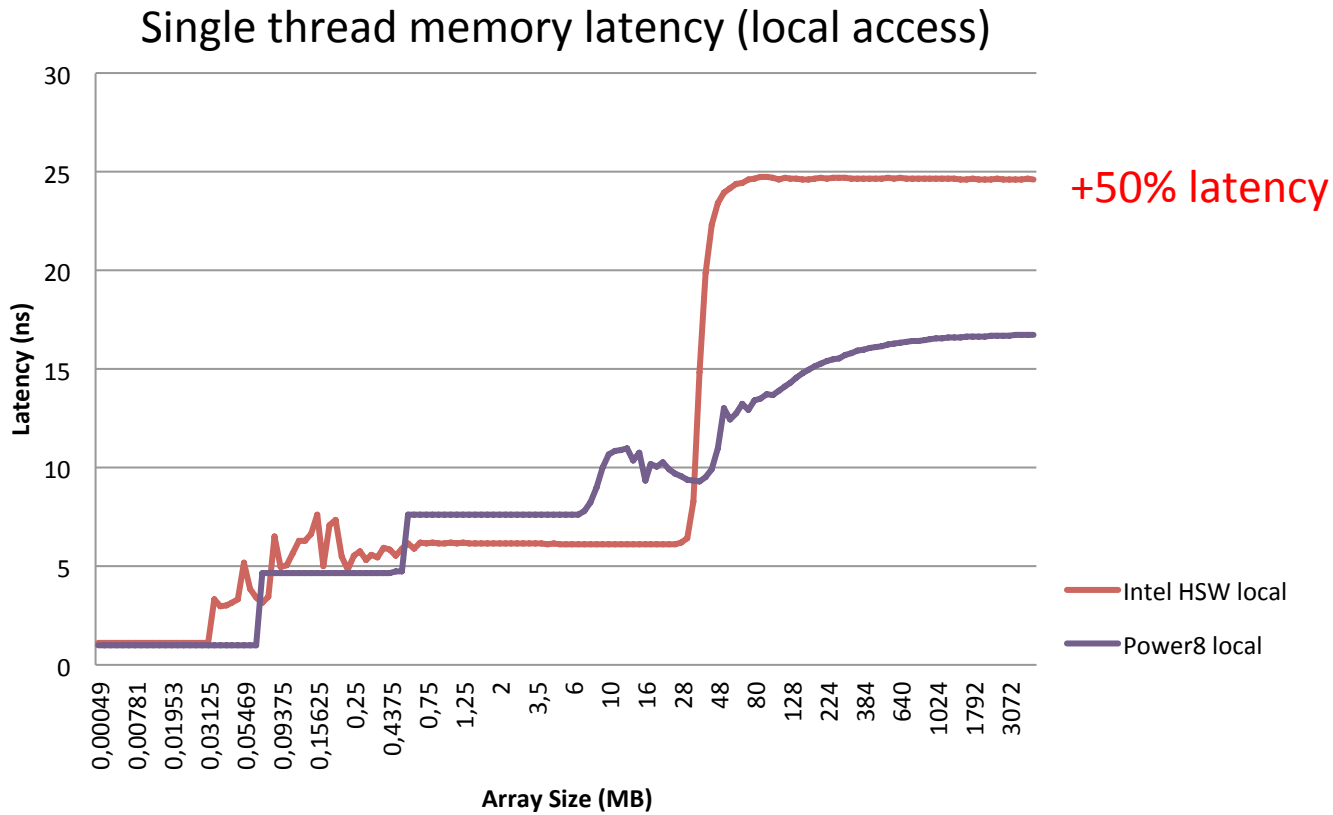
IBM S822 with 2 POWER8
(10 core, 3.42 GHz) **DDR3/1600**

VS.

Dell R630 with 2 Xeon E5-2660 v3
(10 core, 2.6 GHz) **DDR4/2133**

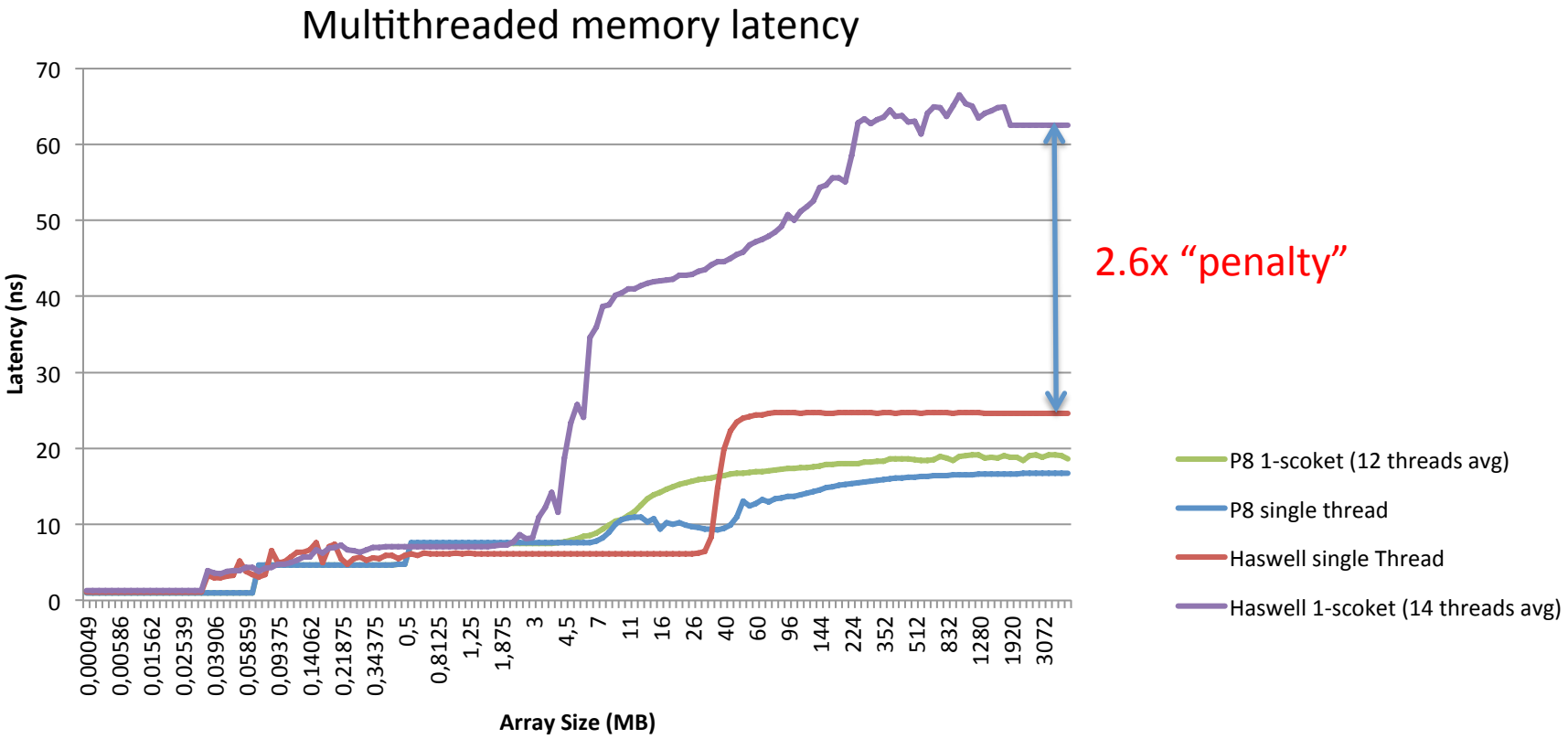
x86 results are from: https://www.cs.virginia.edu/stream/stream_mail/2015/0009.html

Memory latency



Test performed with lmbench comparing S824 (2 socket - 12 cores) and Intel Haswell system (2 sockets – 16 cores) fully populated in memory, Courtesy of Sebastien Chabrolles, IBM France

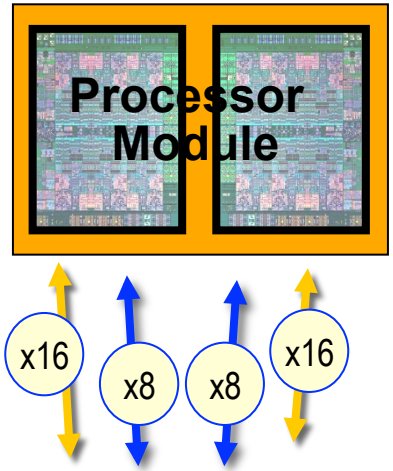
Memory latency + multithreading



Test performed with lmbench comparing S824 (2 socket - 12 cores) and Intel Haswell system (2 sockets – 16 cores) fully populated in memory, Courtesy of Sebastien Chabrolles, IBM France

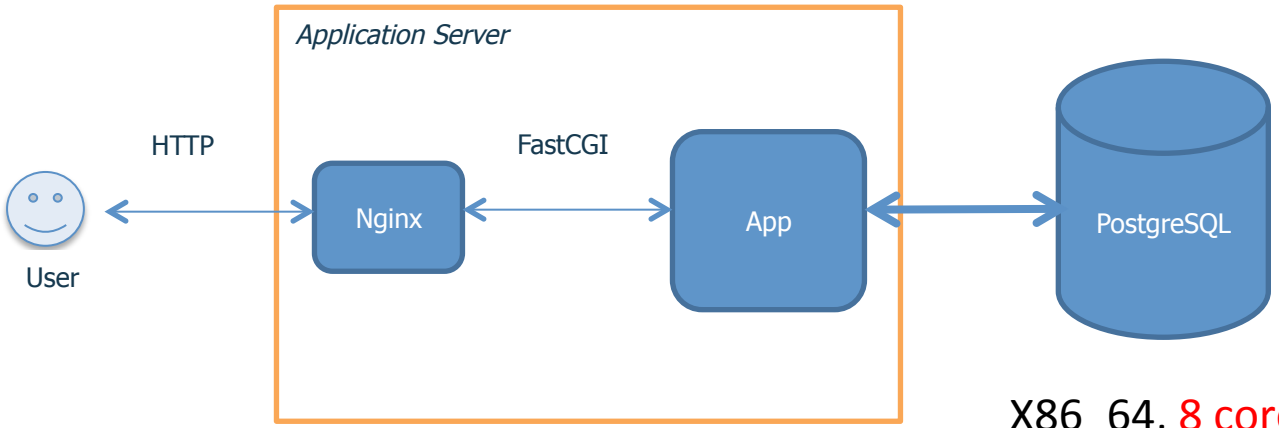
IO performance

1 socket = 1 module



1 socket max bandwidth	96 GB/s
2 socket max bandwidth	192 GB/s

Real world application test



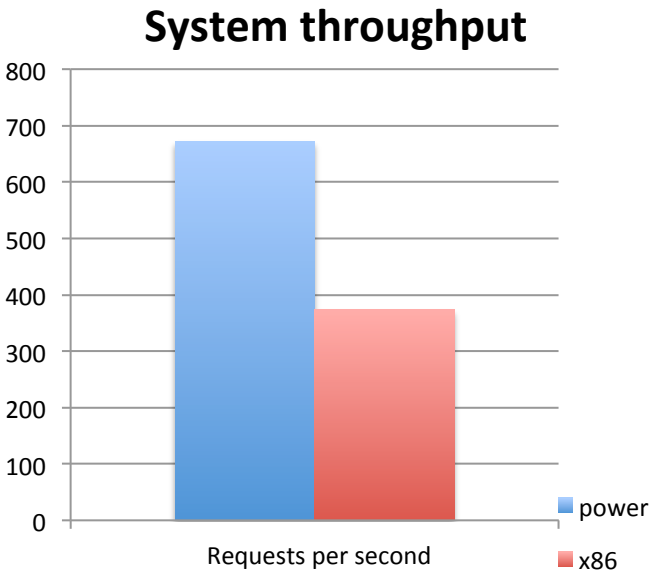
X86_64, 6 cores, 8GB RAM

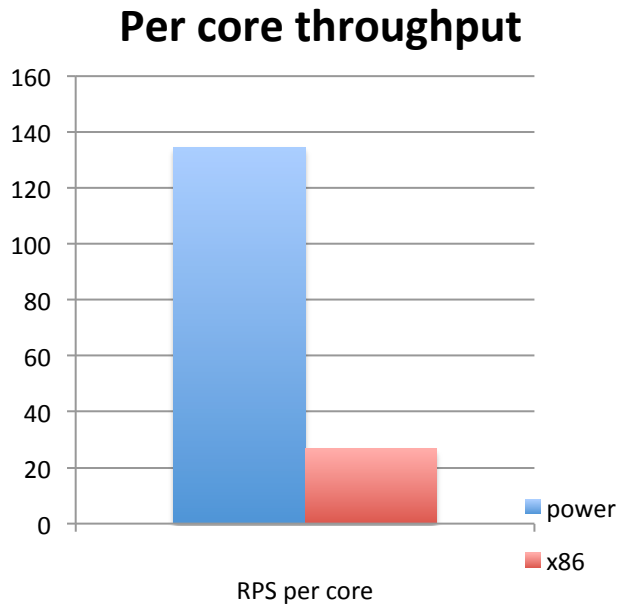
Power8, 3 cores, 8GB RAM

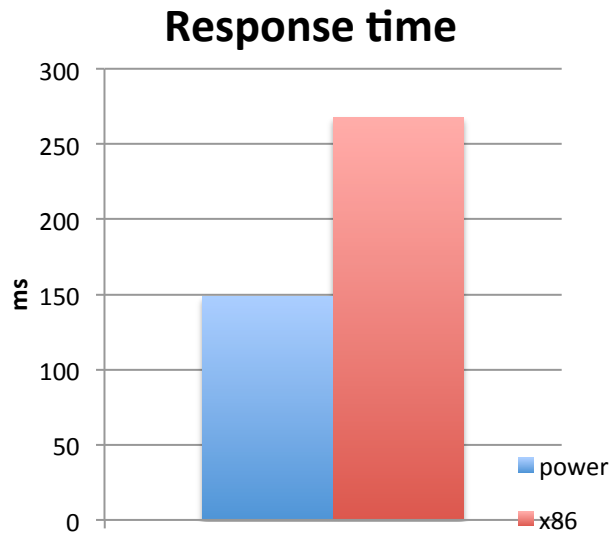
X86_64, 8 cores, 16GB RAM

Power8, 2 cores, 16GB RAM

Real world application test

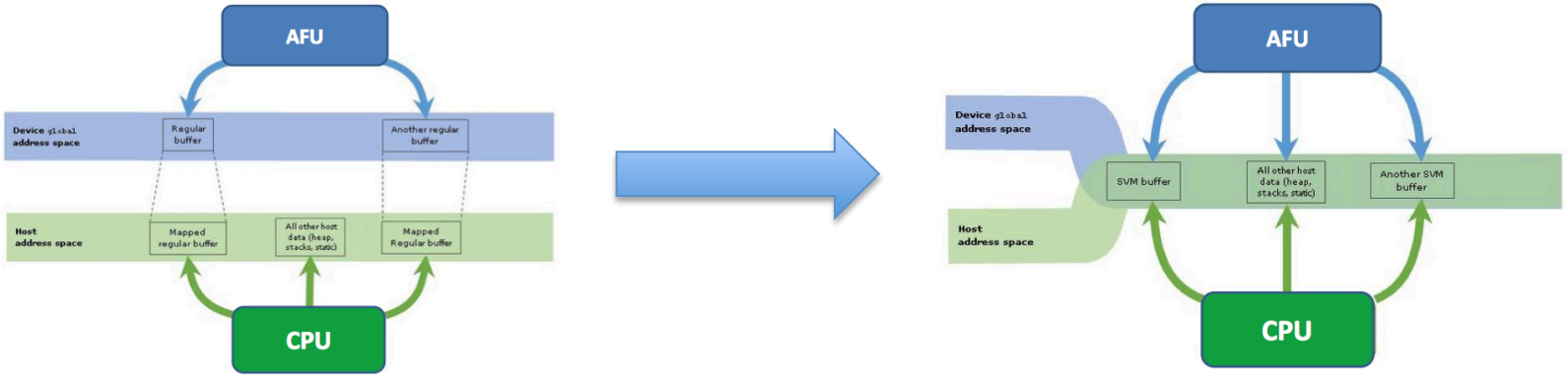




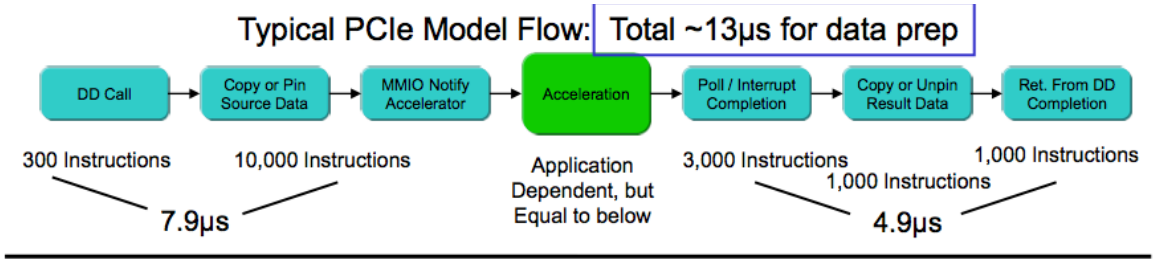


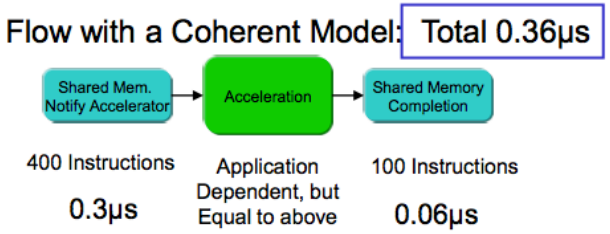
Features unique for POWER8

CAPI – Coherent Acceleration Processor Interface



- Faster than PCIe
- Easy to code

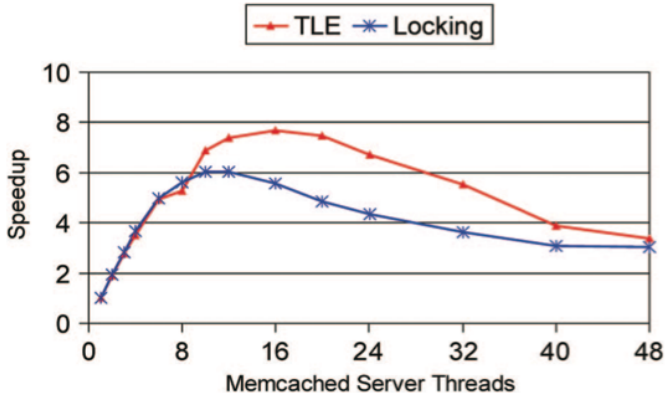


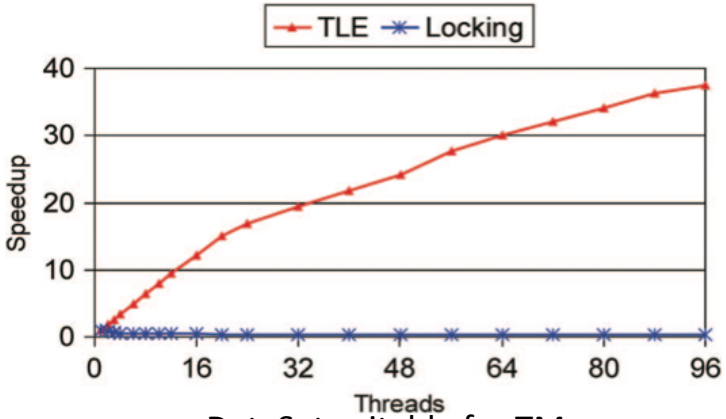


REDIS, **Neo4j** using CAPI
To extend memory to flash.

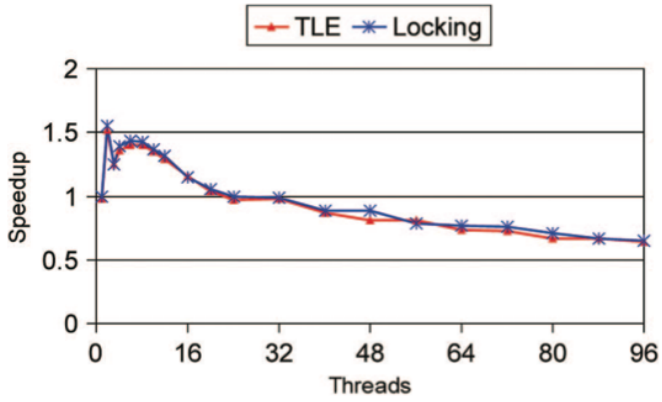
Hardware Transactional Memory

Allow a group of Load/Store instructions execute in atomic way.





DataSet suitable for TM



DataSet overflows TM footprint capacity

Conclusions

- **Server performance** for PG now is more important than ever before.
- Power8 servers are not some transcendental “mainframes”.
- From systems management point of view it is just average server with Linux distro of your choice.
- From performance point of view, Power HW offers best in industry **CPU performance, multithreading, SMP/RAM/IO bandwidths** which makes it ideal choice for PostgreSQL Data Base.
- Unique features that could potentially give even more performance.

Thank you!