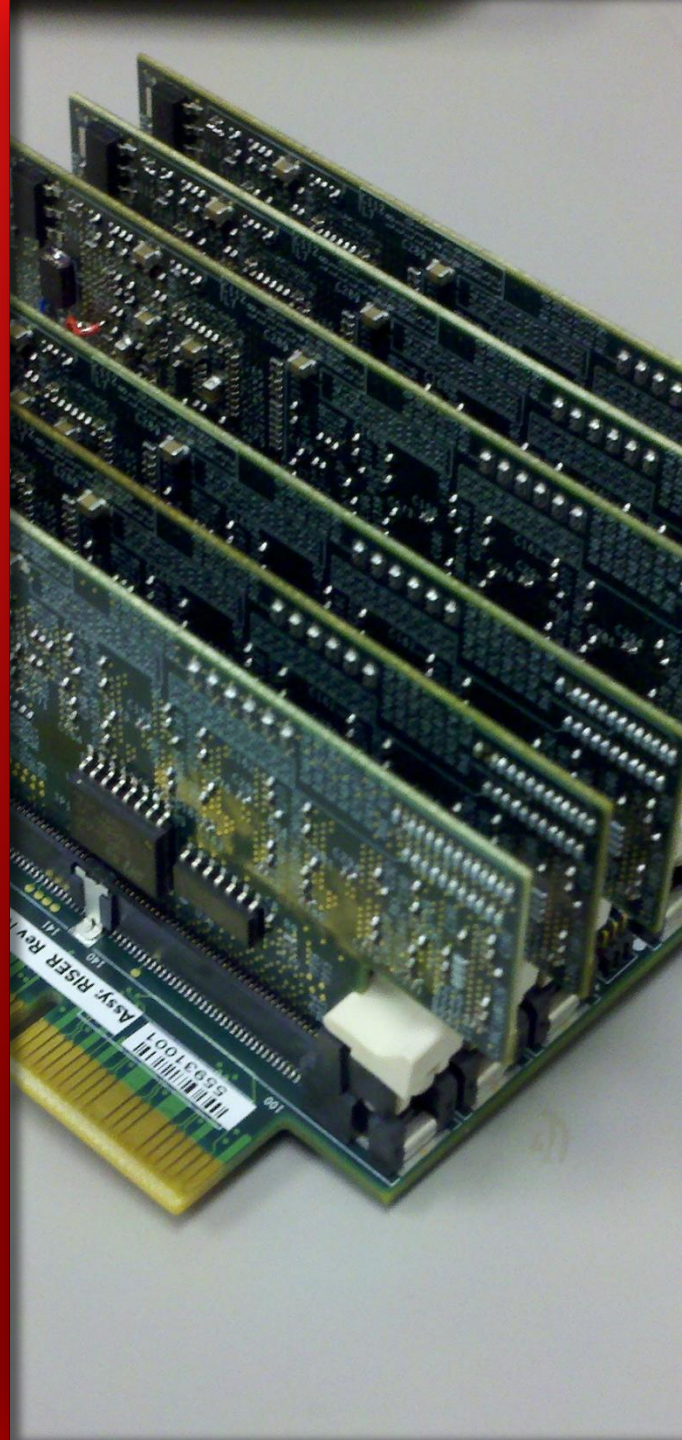
ORACLE

Oracle Labs View on FPGAs

Eric Sedlar
Vice President & Technical Director
Oracle Labs

February 26, 2014



The following is intended to provide some insight into a line of research in Oracle Labs. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. Oracle reserves the right to alter its development plans and practices at any time, and the development, release, and timing of any features or functionality described in connection with any Oracle product or service remains at the sole discretion of Oracle. Any views expressed in this presentation are my own and do not necessarily reflect the views of Oracle.

A Brief History of FPGAs at Oracle / Sun

- Brief investigation into FPGAs as a part of comparative analysis vs Neteeza
  - FPGA accelerators at the disk for high-efficiency scans
  - Oracle conclusion: lower-end X86 processors just as good
    - No new tool chain issues—no need for Verilog expertise
- Sun Microsystems view on FPGAs: companies that know what they are doing will tape out an ASIC
  - FPGAs are for kids & academics
- Sun Labs FPGA-based massively parallel simulator project “Phaser”
  - Fairly successful from the Sun Labs point of view
  - Tool chain incompatibilities

Trends in Enterprise Computing

Engineered systems for rack-level integration

- Nobody buys a single server
- Significant difficulties in configuration of high-end rack-scale systems
  - Infiniband
  - System balance
    - How much flash? How much disk?
  - Many patches to be applied
- Minimizes the need for on-die integration
  - Separate the workload to heterogeneous components in the rack
  - Provide workload separation onto optimum components where latency is not critical

Diversity of Machines

Blades have 100+ h/w threads, large machines have 1000s

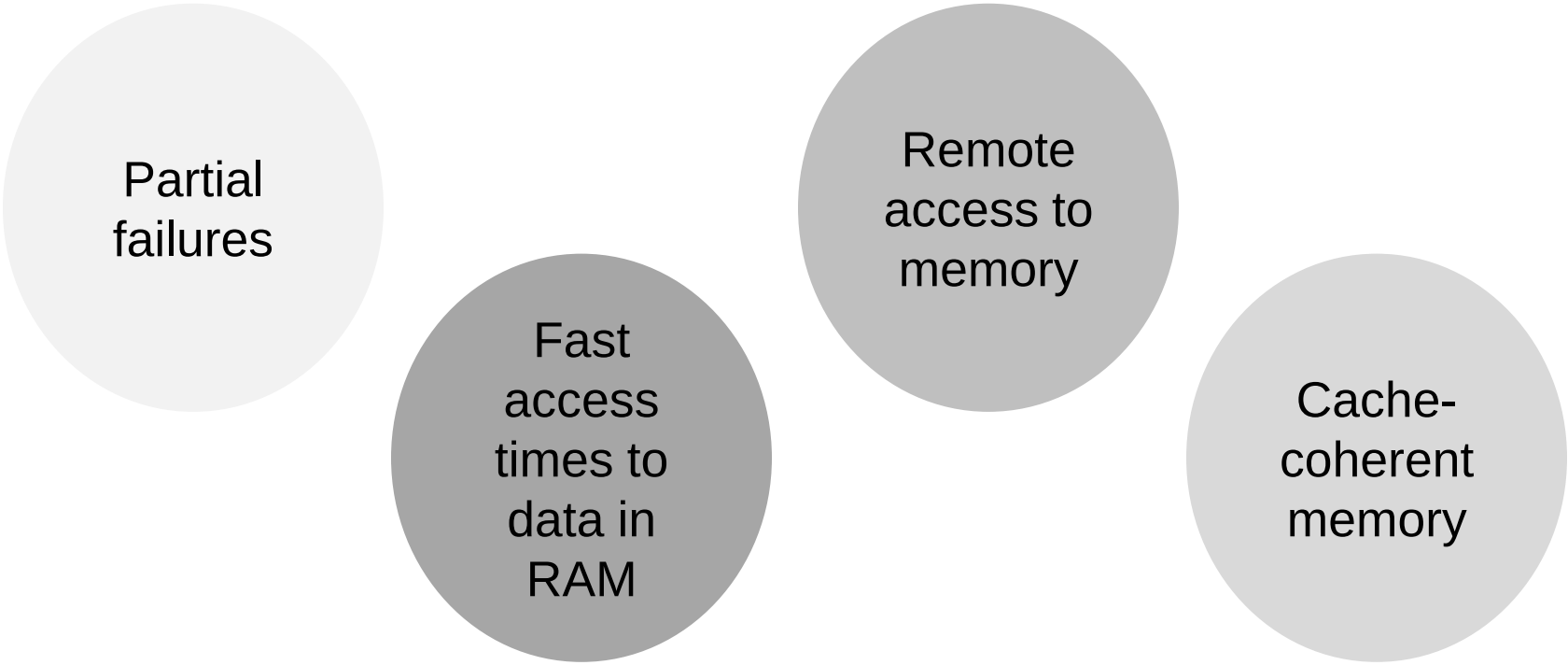
T5-1B
16-cores
128GB-512GB DRAM

SuperCluster T5-8
2 * T5-8 compute nodes
QDR (40 Gb/sec) InfiniBand

SuperCluster M6-32
Up to 32 M6 processors
Up to 32 TB
Cache coherent interconnect

Diversity in Architecture

Boundary becoming blurred between “machine” and “cluster”



Infiniband remote memory access only 4X main memory access

Customer interest in engineered systems

- “Private Cloud in a Box”
  - Very low maintenance costs
- Oracle Exadata database appliance
  - Full rack is 8x 2-socket compute nodes + 14 storage nodes
  - “Smart scan”
  - Multi-billion dollar business
  - 10X performance improvement for most customers
- Exalytics, Exalogic, BDA

Oracle move to “Software in Silicon”

You keep using that word…I do not think it means what you think it does

- In a rack-engineered system, many opportunities for acceleration
- SPARC M7+ will start seeing features to support database & Java migrate into accelerators in the processor & network hardware
- Captive software base to accelerate
- This work is all being done with an ASIC design flow
  - Very expensive and difficult

Good opportunity for using FPGAs

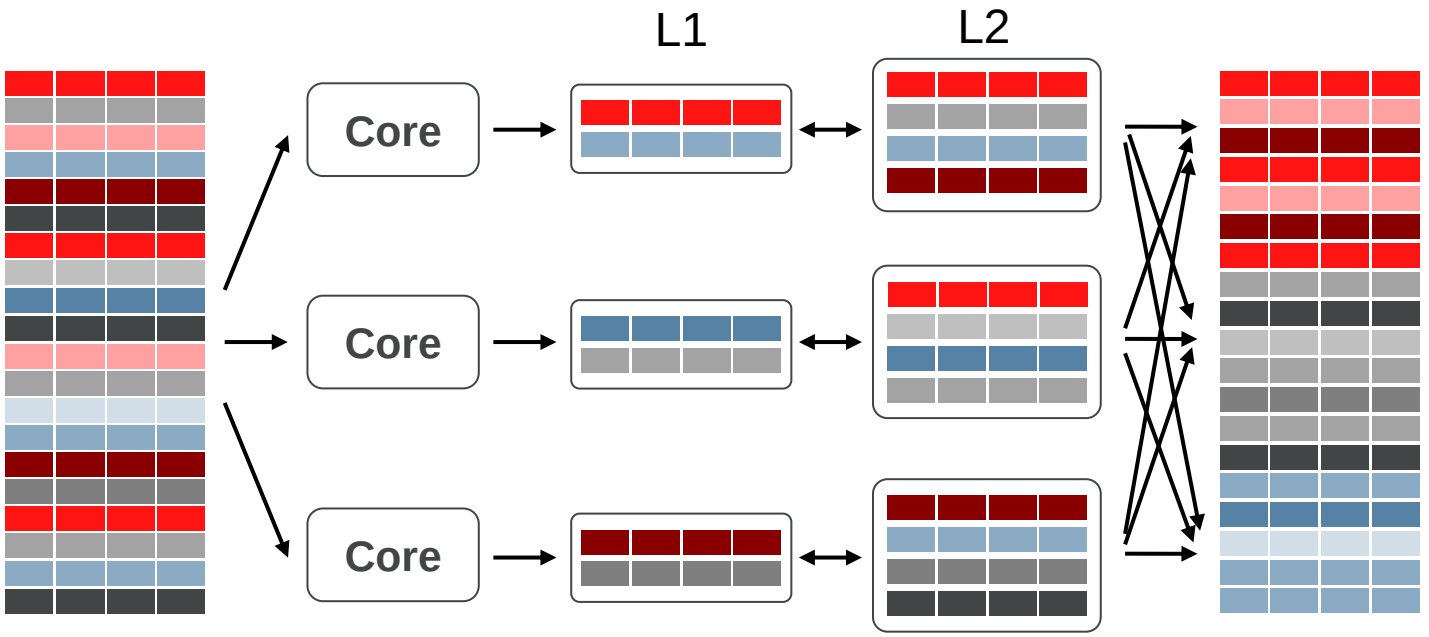
- FPGAs have much lower development cost when you start with a bunch of pre-existing C code and you want to build an accelerator
  - Existing code has unit tests that help tremendously with verification
  - Easier to have a software-style development cycle (debug things into existence) which helps as you incrementally decide how much software to offload
- Offload has a tremendous maintenance cost
- Need to validate that each chunk of software offloaded is worth it

Co-design Example: Group-By & Aggregation

- Performance constraints vary with query, dataset, and algorithm
  - Case 1: DRAM bandwidth-bound
    - Small cardinality, simple key types, small number of aggregates
  - Case 2: Cache- sensitive (size, latency, parallelism)
    - Medium cardinality, simple key types
  - Case 3: Instruction throughput-bound
    - Medium cardinality, variable-length keys or complex aggregates
  - Case 4: Large # of partitions blow up TLB & cache streaming prefetch algorithms
    - For large cardinality, you want to partition input to chunks that fit in cache
- To achieve significant speedups across this range, we need
  - **Flexible acceleration hardware:** programmable cores w/ specialization
  - **Balanced design:** trade off hardware features with power & design complexity

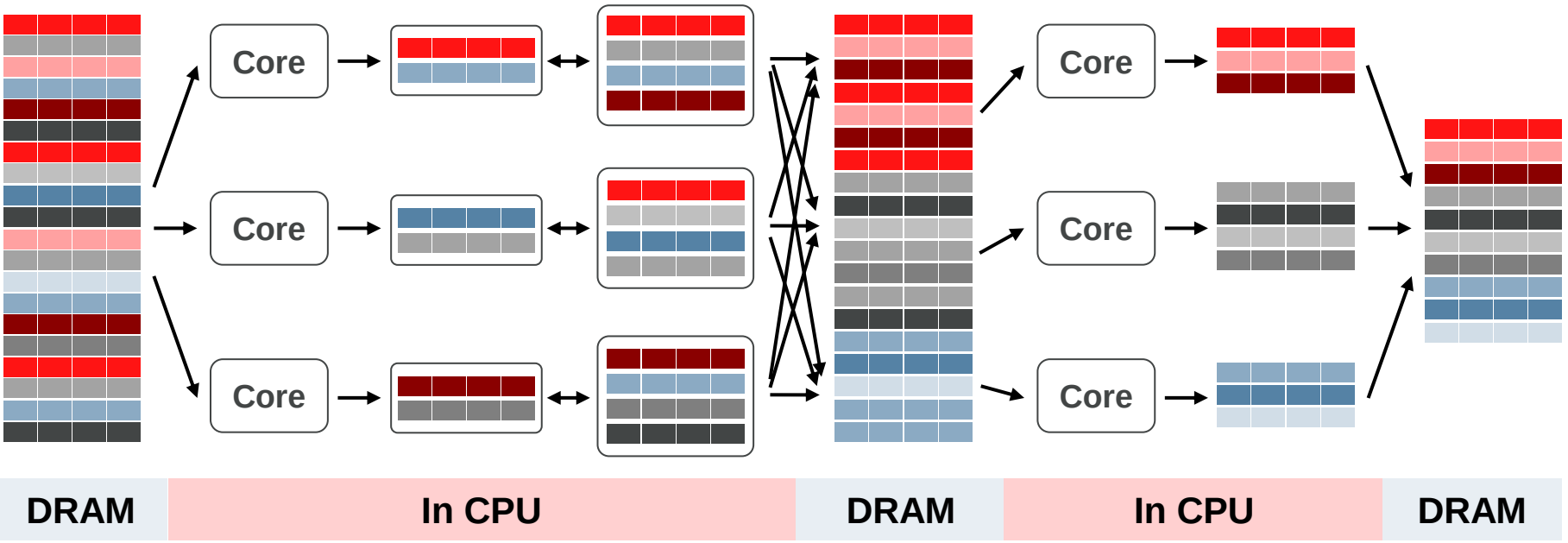
By doing this analysis across different workloads we can identify common HW features and requirements

Partitioning in Conventional CPUs



- Use caches to buffer random writes to DRAM
- More partitions require larger caches for more buffering
  - Need high-end, OoO core to hide L2 hit latency ➡ **power**
  - Need multi-level caches ➡ extra **power** in L1/L2 data movement
- State-of-the-art on x86 server: 1024-way partitioning at DRAM B/W

CPUs Move Data Excessively



- Extra power, performance overhead in data movement
  - Multiple passes over data at high bandwidth
  - Cache lines ping-ponging between levels at high speed

Can we do better?

FPGA projects in Oracle Labs

- KeyBridge: Application logic in the NIC
  - Programmable “Fast Path” that runs at line rate with a DSL
    - Not a Turing-complete machine
  - Integrated with embedded Java
  - Logging to an Oracle database
  - Markets: Financials, Telco, Cybersecurity
- Database FastLoader
  - Compression & block formatting slow down database loading to 10% of disk BW or network BW when using all cores
  - A better answer to “NoDB”
- How can we manage skew in our rack-integrated cluster?
  - Need management tools in the fabric

The ecstasy & agony of HW / SW co-design

Need very good end-to-end system workloads

- Load Rates - Direct Path (existing software):

Exadata X2-2 Half Rack 1.44TB data, Query High, 90 Parallel x 4 compute nodes.

Data Type	Load Rate	Storage GB/s [1]	**Offloadable [2]**
Number	2.09 GB/sec	0.21 - 0.52	**> ~78-93%**
Varchar2(39)	1.39 GB/sec	0.14 - 0.35	**> ~84-95%**

Notes: (1) Measured storage rate, then scaled for 10x to 4x compression measured in benchmarks.

(2) Preliminary estimate, from code profile.

- Disk and I/O Channel Headroom :

Write rate with normal redundancy	HighCap	HighPerf [3]
Compression=none (steady-state)	4.5 GB/s	6.2 GB/s
Headroom (underutilization)	9x	12x

FPGAs looking like a good fit at Oracle

- Lots of low volume high dollar verticals
  - Don’t need many customers to recoup engineering investment
  - Looking for a performance edge to gain a jump on competition
- High-end customers willing to try cutting edge technology
  - Generally highly educated and technically savvy
  - Good source of vertical requirements
- Easy option for an rack-level engineered system
  - Integration of the FPGA accelerator into an engineered system provides upsell & cross-sell opportunities

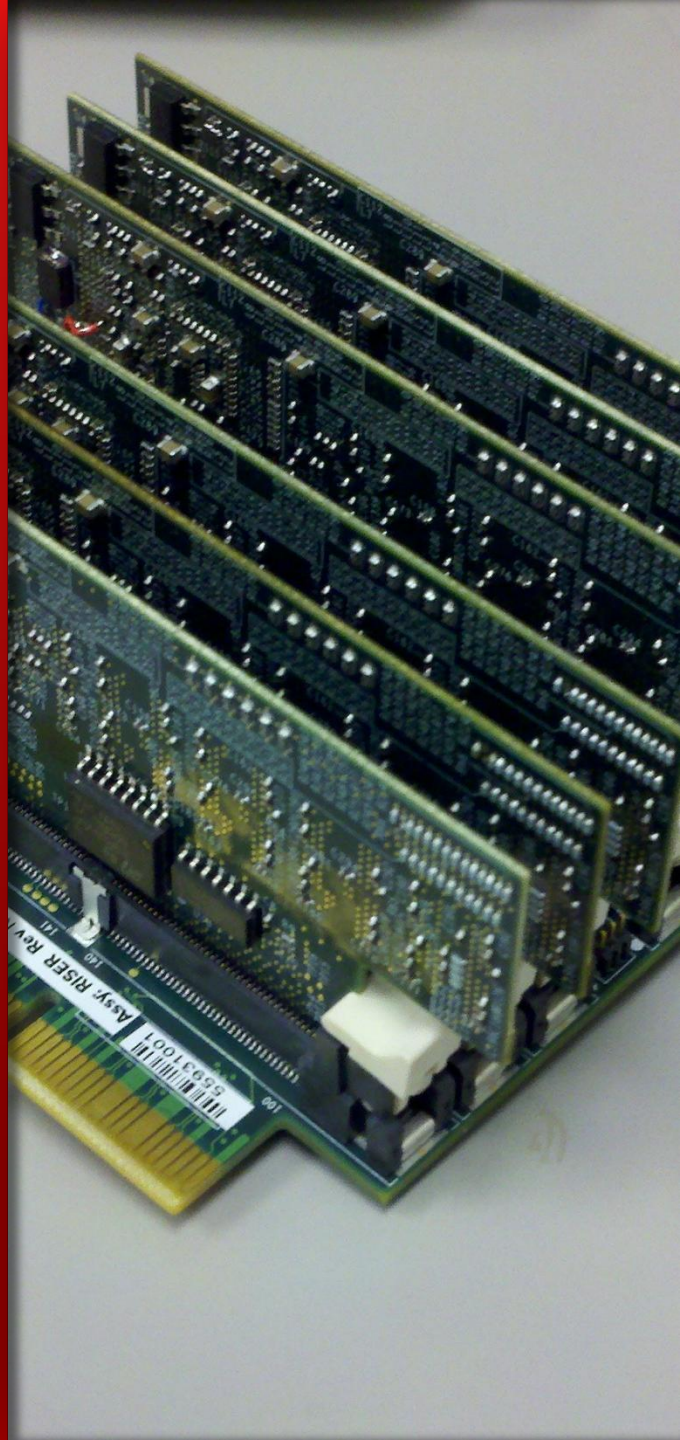
Personal Opinions

- FPGA opportunities are NOT going to be close to the compute engine but rather in the cluster
  - Relatively expensive to go wake up a thread on another node
- Centralized services at the rack level
  - Routing
  - Rack-level management (repartitioning, load balancing, etc) will bottleneck traditional processors
- Big issues:
  - How do I get Infiniband integration on my FPGA card?
    - FDR today (56 GB / port)
    - EDR next year (100 GB / port)

ORACLE

Please visit

http://labs.oracle.com



Hardware and Software

ORACLE®

Engineered to Work Together