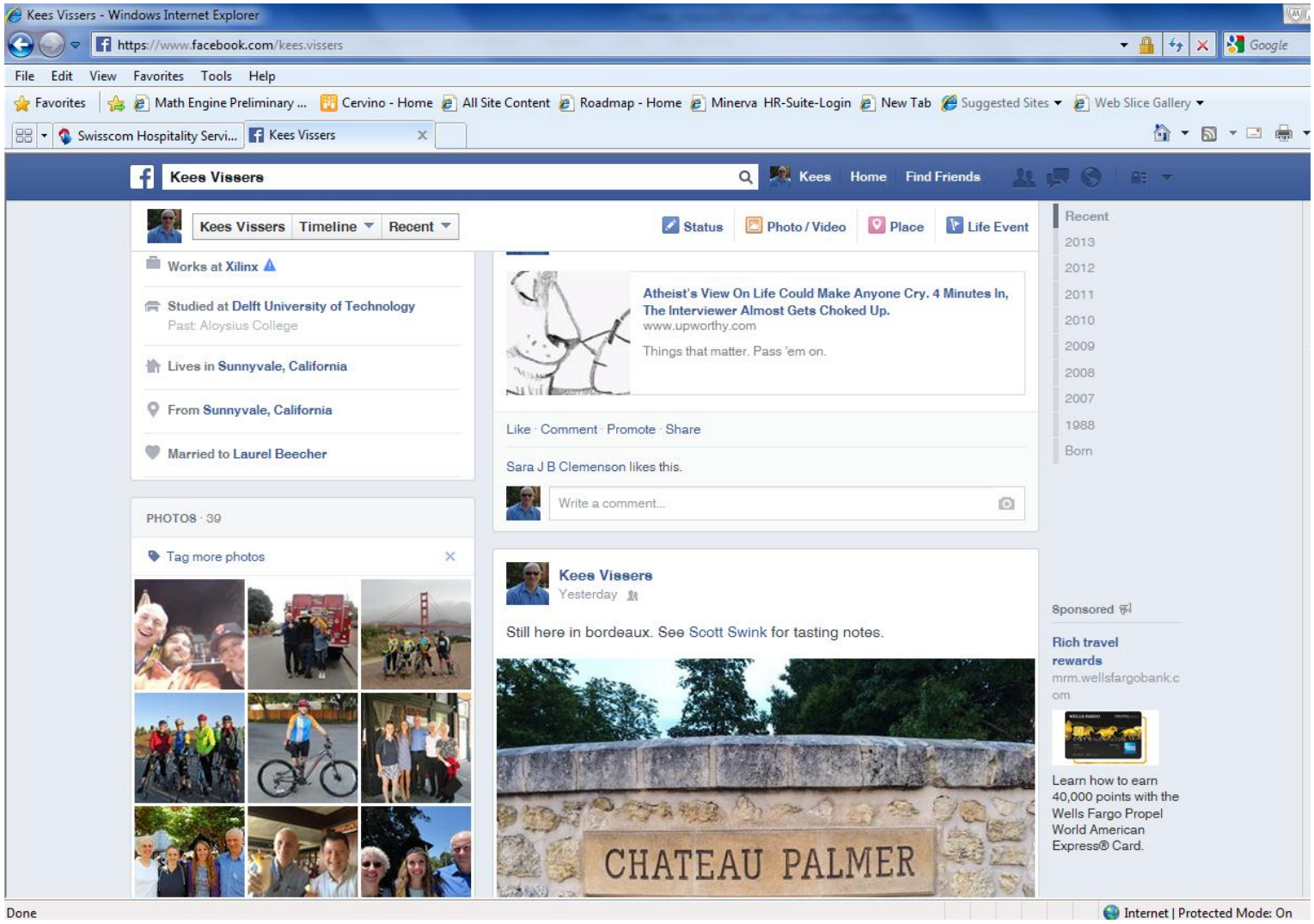
Novel 10Gps Key-Value Stores with High Level Synthesis in FPGAs

Kees Vissers
Xilinx

Facebook



Key-Value Stores

- **Common middleware application to alleviate access bottlenecks on databases**
  - Most popular and most recent database contents are cached in main memory of a tier of server platforms

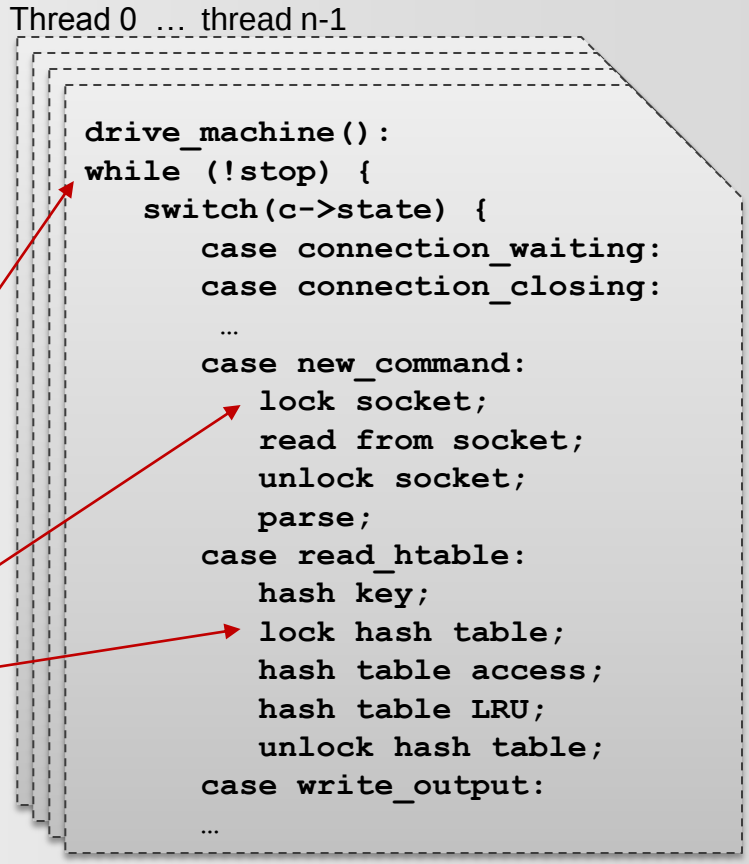
- **Used by many well-known websites**
  - up to 30% of servers in data centers run memcached or similar

- **Current server-based implementations are limited and cannot keep up with 10Gbs network speed**

- **Investigated using dataflow architectures on FPGAs to dramatically increase performance and lower power and latency**

Typical Implementations

- **Software**
  - Each connection is represented as a struct (c)
  - Any event on the connection state is distributed to pthreads (via Libevent)
  - All worker threads run the same code (drive_machine())
    - Loop over switch statement over the connection state
    - Locks on sockets, hash table, and value store areas/items



Bottlenecks

- **TCP/IP**
  - CPU intensive (114% system cycles vs 45% user space out of 800%)
  - Large footprint
    - Leads to high rate of instruction cache misses (up to 160 MPKI)
  - Frequent interrupts
    - Causes poor branch predictability (stalling superscalar pipeline) on x86

- **Synchronization overhead**
  - Threads stall on memory locks, serializing execution for x86s
- **Last level cache ineffective due to random-access nature of the application (miss rate 60% - 95% on x86)**
  - Multithreading can't effectively hide memory access latencies
  - Causes considerable power waste
- **High latency**
  - Packets have to be DMA'ed from/to network adapter over the PCIe® bus which introduces high latency

Best Published Performance Numbers

1.4MRPS

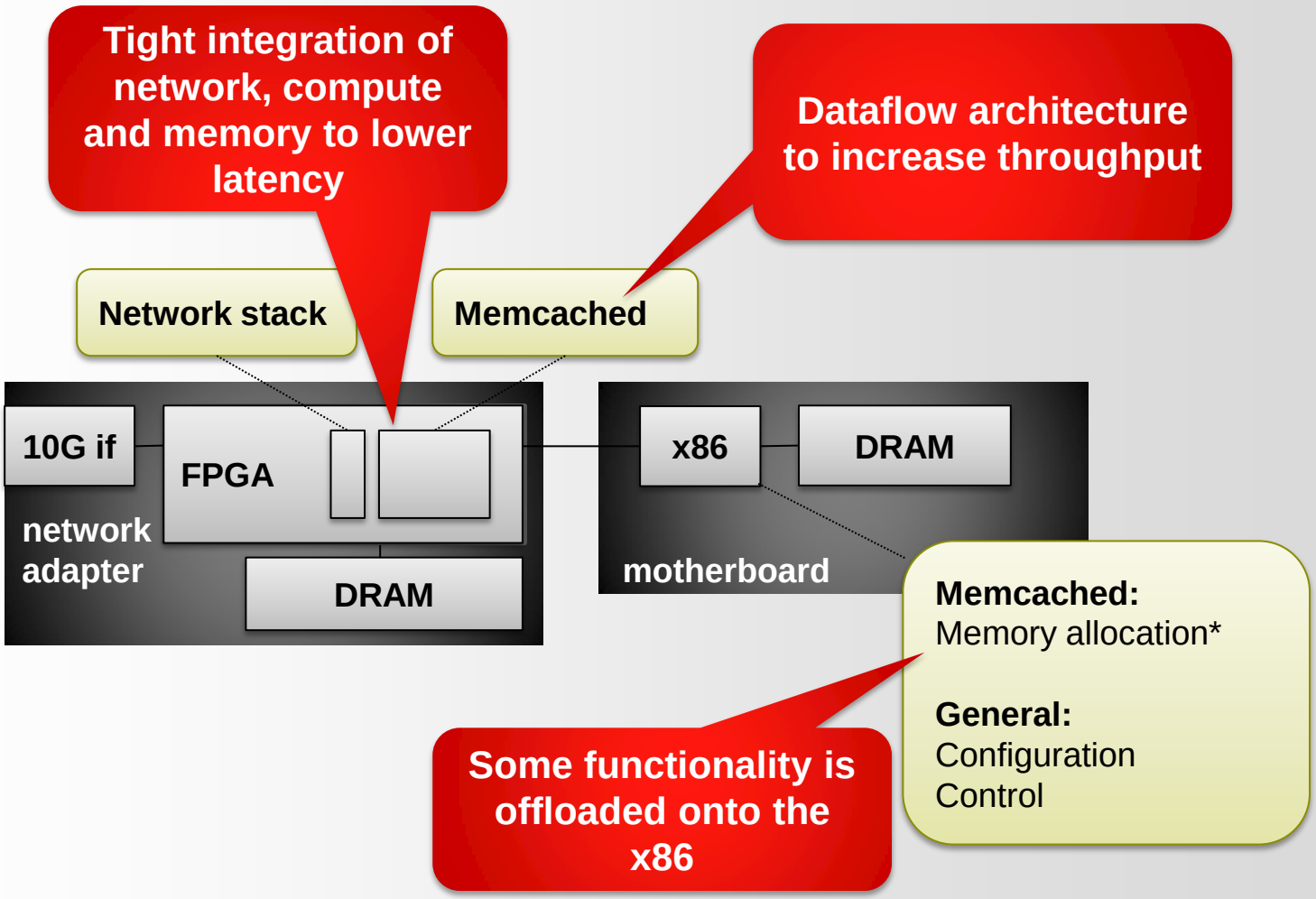
200us latency

7K

Platform	RPS [M]	Latency [us]	RPS/W [K]
Intel® Xeon® (8 cores)*	1.34	200-300	7
Memcached with Infiniband & Intel Xeon (2 sockets, 16cores)**	1.8	12	Unknown
TilePRO (64 cores)***	0.34	200-400	3.6
TilePRO (4x64 cores)***	1.34	200-400	5.8
Chalamalasetti (FPGA)****	0.27	2.4-12	30.04

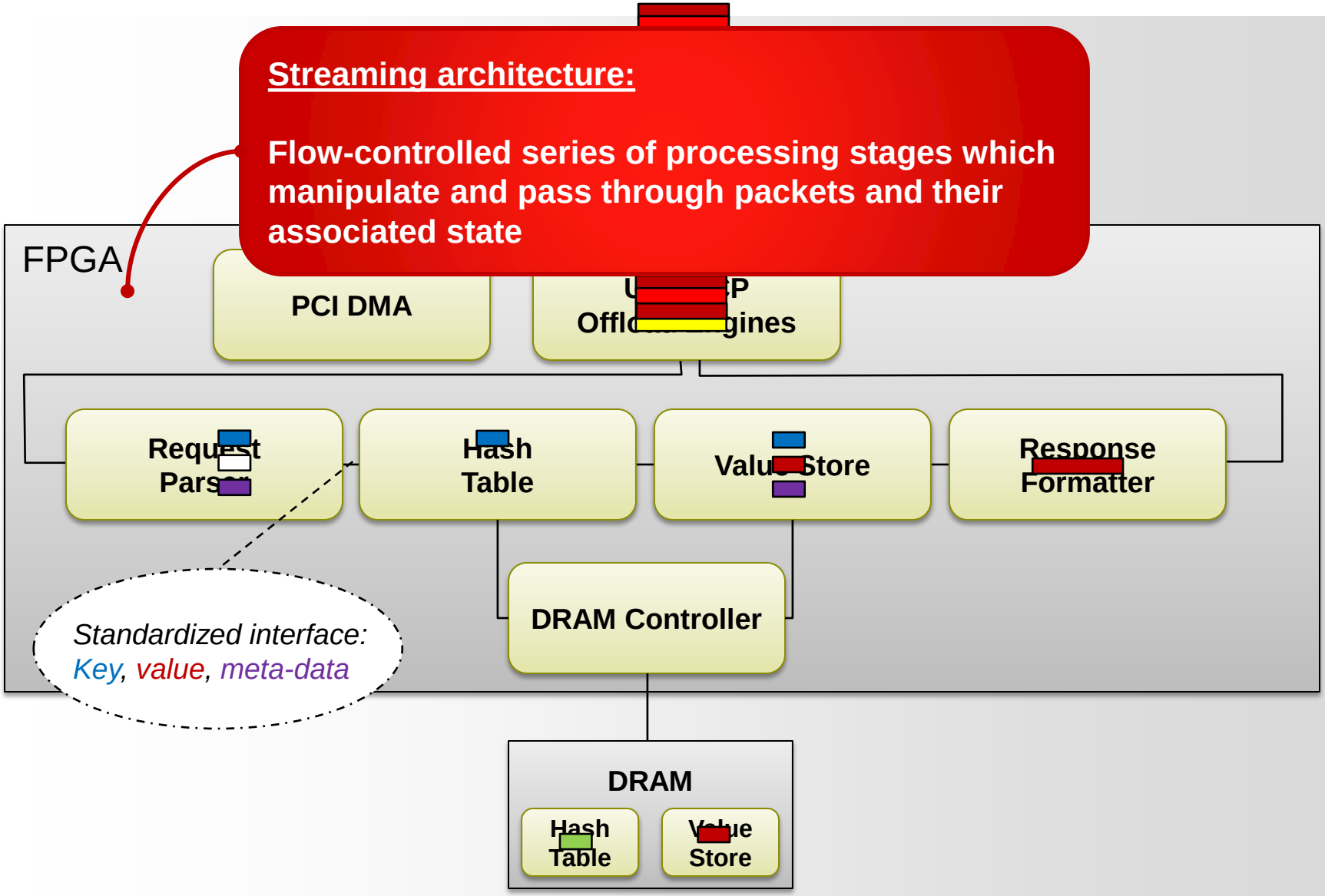
* WIGGINS, A., AND LANGSTON, J. Enhancing the scalability of memcached. In Intel Software Network (2012).
**JOSE, J., SUBRAMONI, H., LUO, M., ZHANG, M., HUANG, J., UR RAHMAN, M. W., ISLAM, N. S., OUYANG, X., WANG, H., SUR, S., AND PANDA, D. K. Memcached design on high performance rdma capable interconnects. 2012 41st International Conference on Parallel Processing 0 (2011), 743–752.
*** BEREZECKI, M., FRACHTENBERG, E., PALECZNY, M., AND STEELE, K. Power and performance evaluation of memcached on the tilepro64 architecture. In Green Computing Conference and Workshops (IGCC), 2011 International (July 2011), pp. 1 –8.
**** Kevin Lim, David Meisner, Ali G. Saidi, Parthasarathy Ranganathan, and Thomas F. Wenisch. 2013. Thin servers with smart pipes: designing SoC accelerators for memcached. In*Proceedings of the 40th Annual International Symposium on Computer Architecture* (ISCA '13). ACM, New York, NY, USA, 36-47.
**** CHALAMALASETTI, S. R., LIM, K., WRIGHT, M., AUYOUNG, A., RANGANATHAN, P., AND MARGALA, M. An fpga memcached appliance. In Proceedings of the ACM/SIGDA international symposium on Field programmable gate arrays

System Architecture

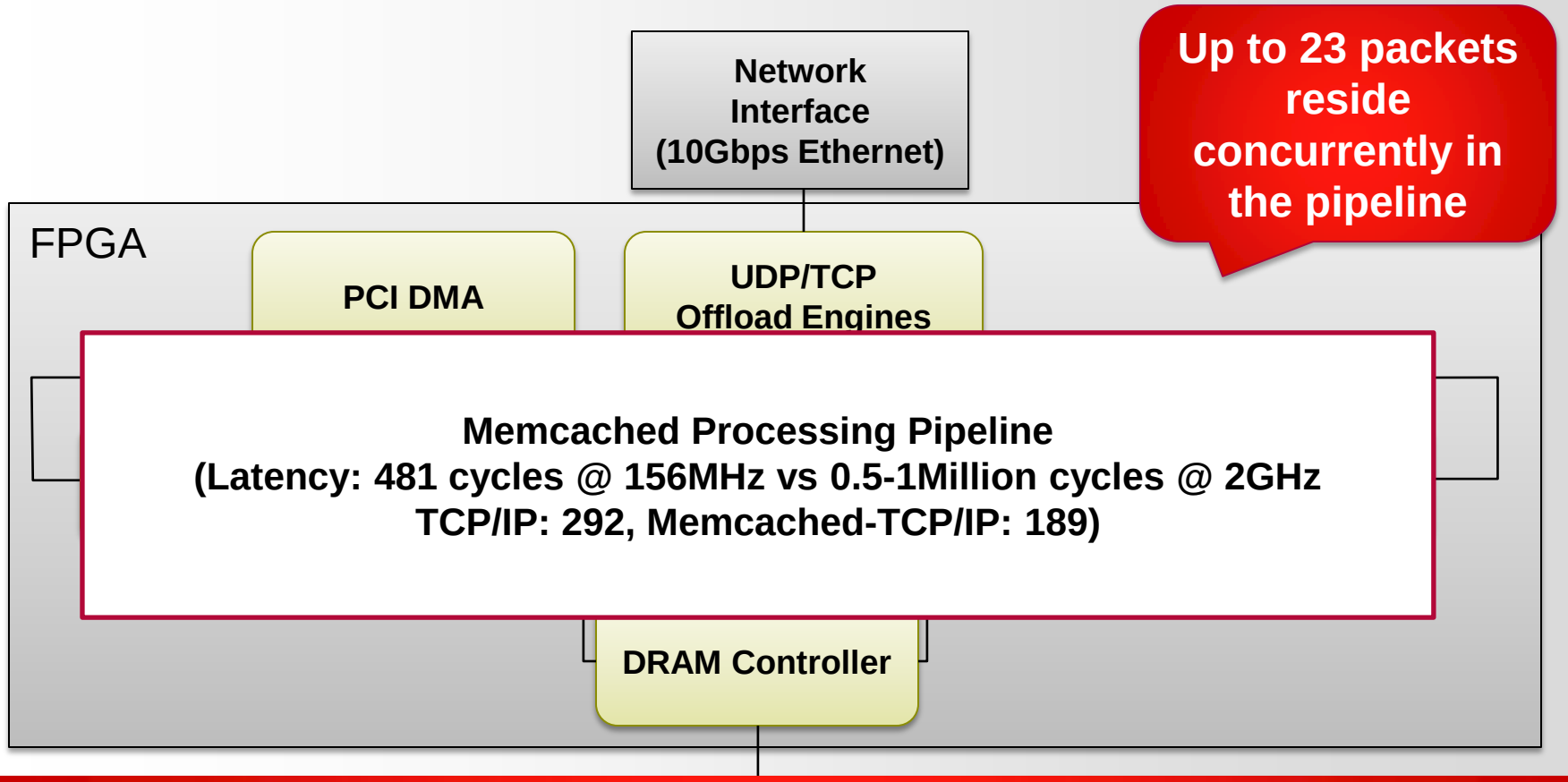


*below 3% of 1 core for 10% SET operations
*limited memory access bandwidth on platform

Dataflow Architecture



Dataflow Architecture

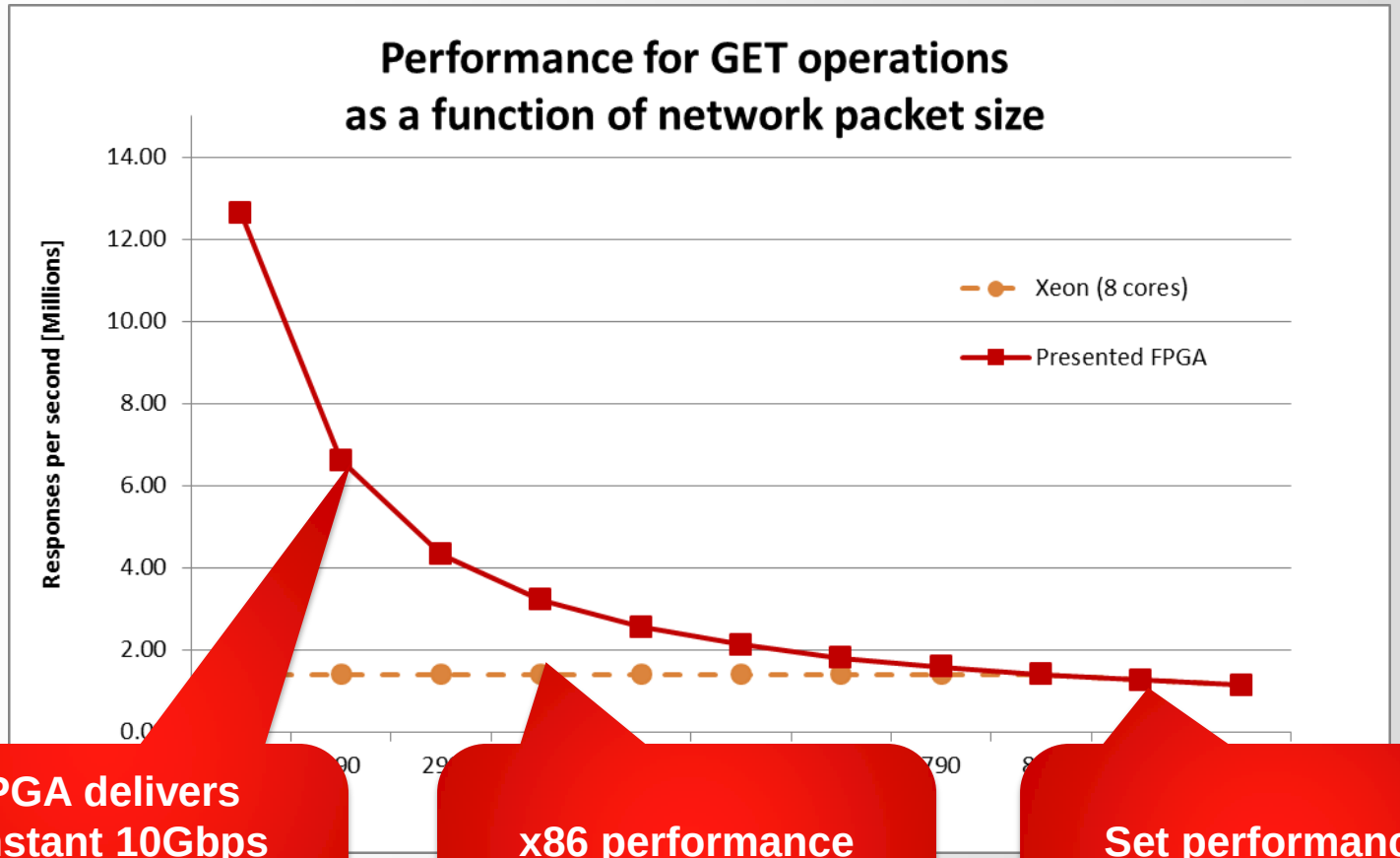


=> Exploiting fine-grain parallelism increases throughput, lowers latency and is more power efficient

=> Inherently scalable

Results - Performance

Higher is better *** UDP, GET, BINARY**



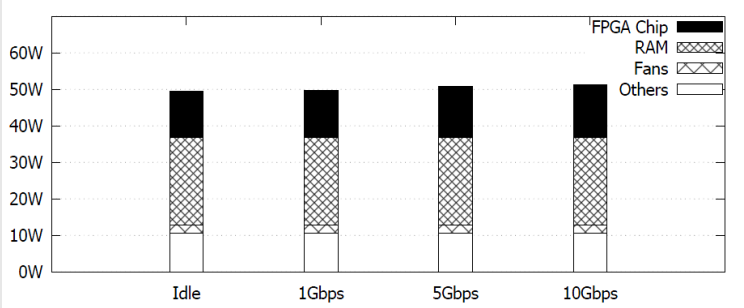
FPGA delivers constant 10Gbps performance – network becomes the bottleneck

x86 performance limited by a per packet overhead

Set performance saturates network as well

Results

- **Sustained line rate processing for 10GE – 13MRPS possible, at smallest packet size**
  - Significant improvement over latest x86 numbers
- **Lower power**
  - Combined: 36x in RPS/Watt with low variation
- **Cutting edge latency**
  - microseconds instead of 100s of microseconds



- **HLS design flow validated for entire memcached functionality**
  - Reducing code and development time by half while bringing down resources

Platform	RPS [M]	Latency [us]	RPS/W [K]
Intel Xeon (8 cores)	1.34	200-300	7
TilePRO (64 cores)	0.34	200-400	3.6
FPGA (board only)	Up to 13.02	3.5-4.5	254.8
FPGA (with host)	Up to 13.02	3.5-4.5	106.7

The FPGA benefit for small value sizes

Calculated probability of value sizes									
Value size [Bytes]	128	256	512	768	1014	2048	4096	22000	32000
Facebook: ETC	0.55	0.075	0.285	0.015	0.025	0.025	0.025	0	0
Facebook: USR	1	0	0	0	0	0	0	0	0
Facebook: APP	0.12	0	0.63	0.21	0.03	0.01	0	0	0
Facebook: VAR	0.78	0.02	0.17	0.03	0	0	0	0	0
Twitter	0	0	0	0.1	0.85	0.05	0	0	0
Wiki	0	0	0	0	0.58	0.02	0.1	0.25	0.05
Flicker	0	0	0	0	0	0	0	0.1	0.9
Youtube	0	0	0	0	0	0.75	0.11	0.11	

metadata, user-account status information, server-side browser information, nonspecific, general-purpose information

thumbnails

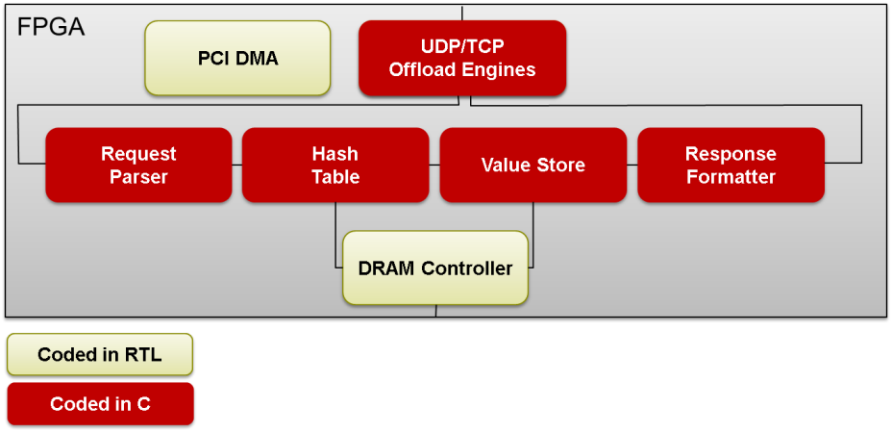
thumbnails

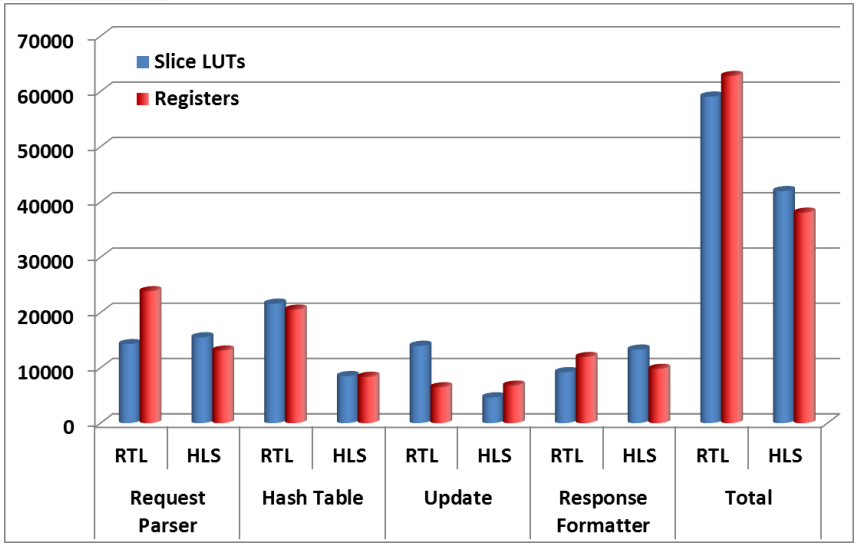
Tweets + meta data

articles

Design Flow with HLS

- Higher-level design flow (Vivado HLS) validated for memcached functionality
- Reduced code and development time by more than half
- All modules meet timing and throughput requirement
- Resource slightly reduced in comparison to RTL





Key Results and Next steps

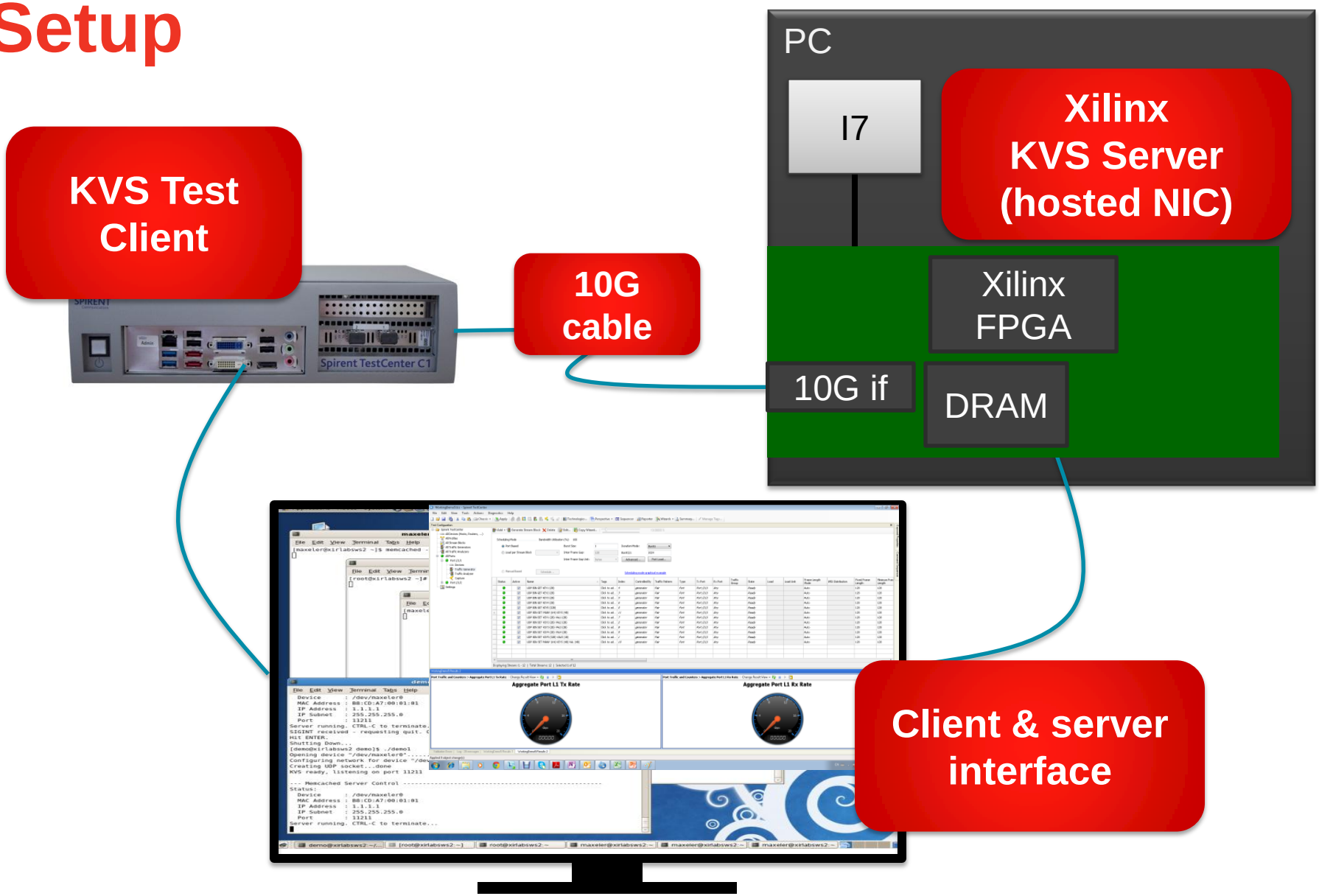
- **Sustained line rate processing for 10GE**
  - 13MRPS at smallest packet size
- **Lower power**
  - 15W FPGA vs 54W CPU
  - Combined: 36x in RPS/Watt with low variation
- **Cutting edge latency**
  - Microseconds instead of 100s of microseconds

Next steps:

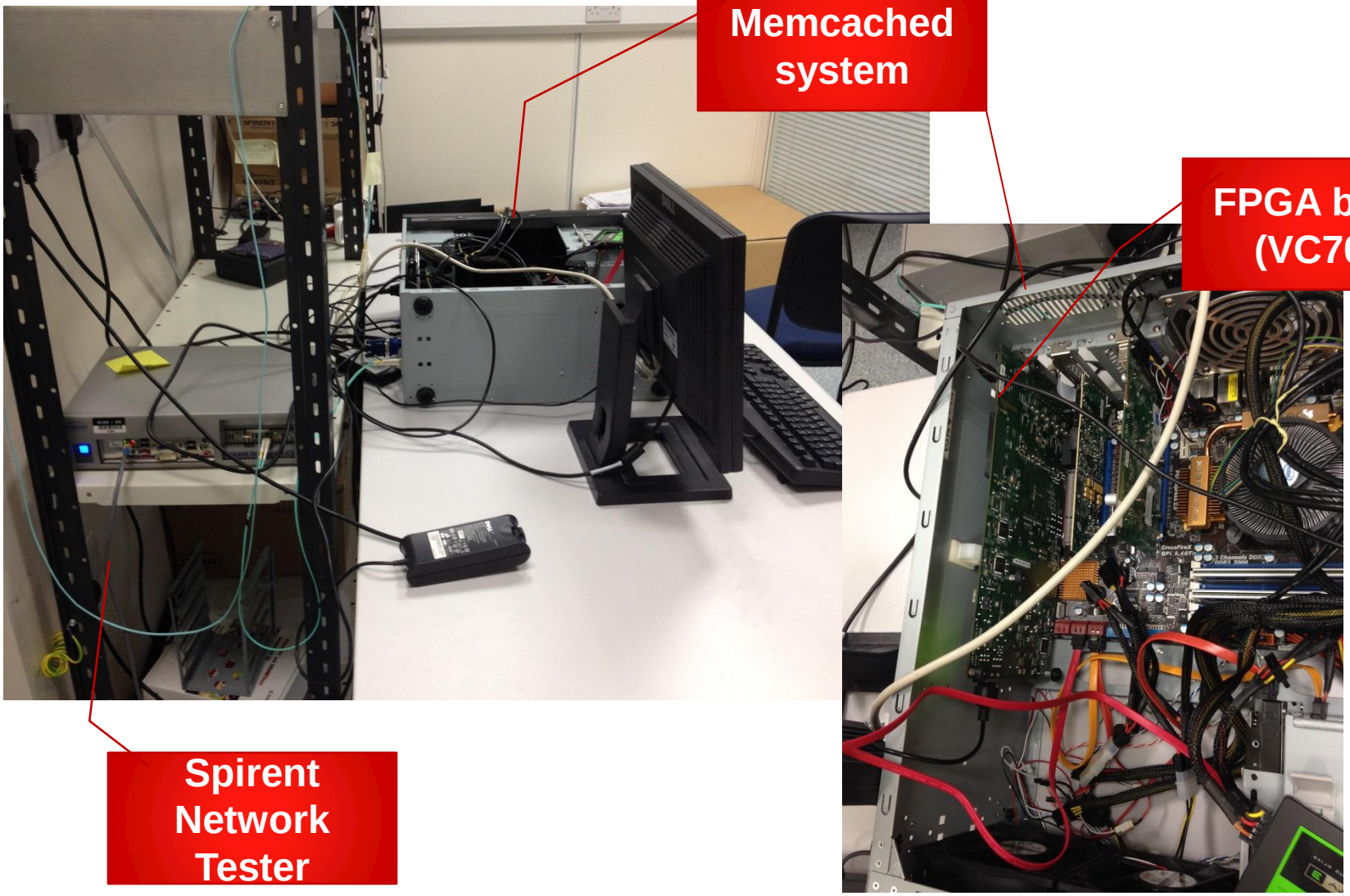
- **Integrate with Flash based technology for large storage**
- **Investigate next generation ARM based FPGA combinations on Xilinx Zynq devices**

The measurement setup

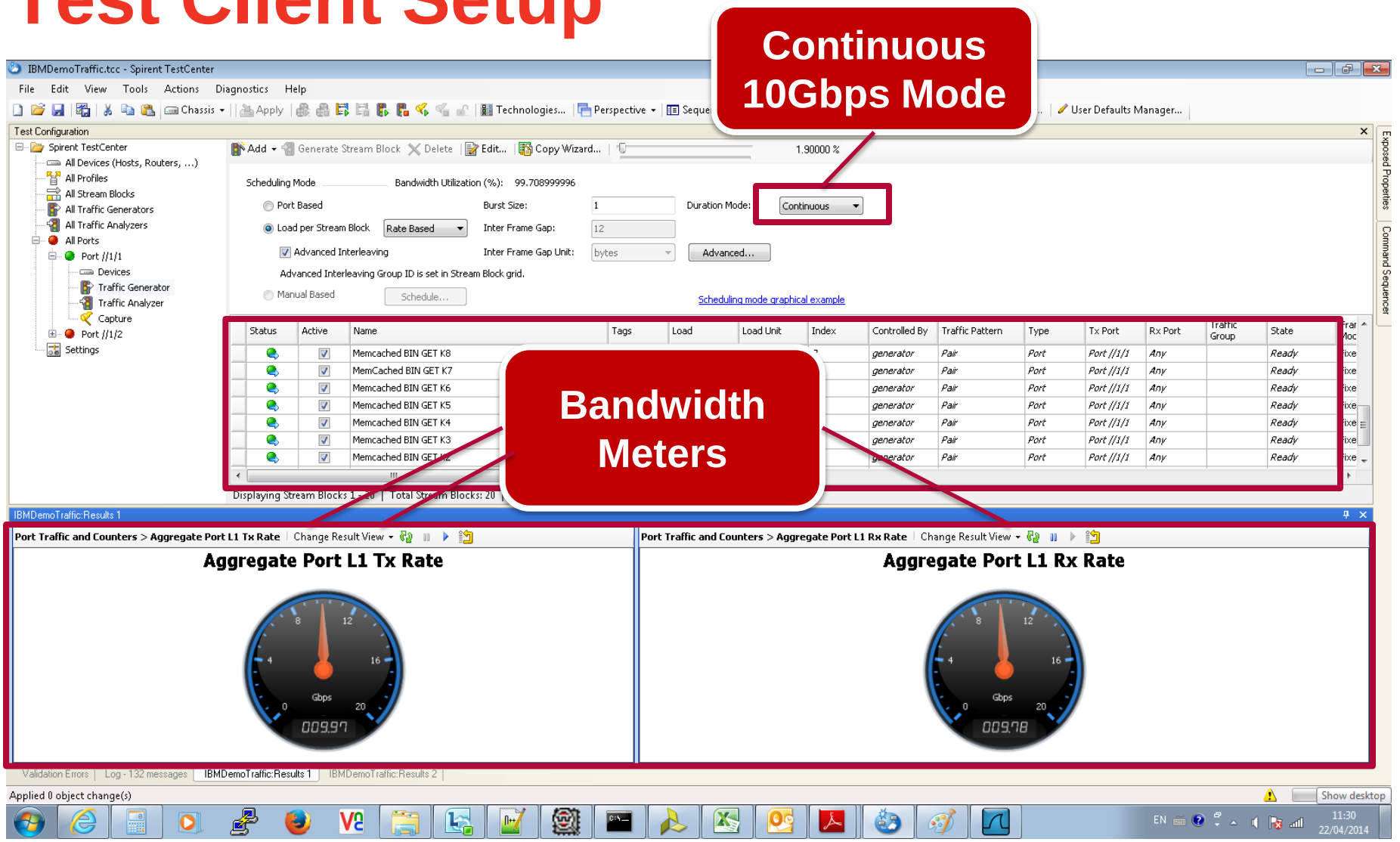
Setup



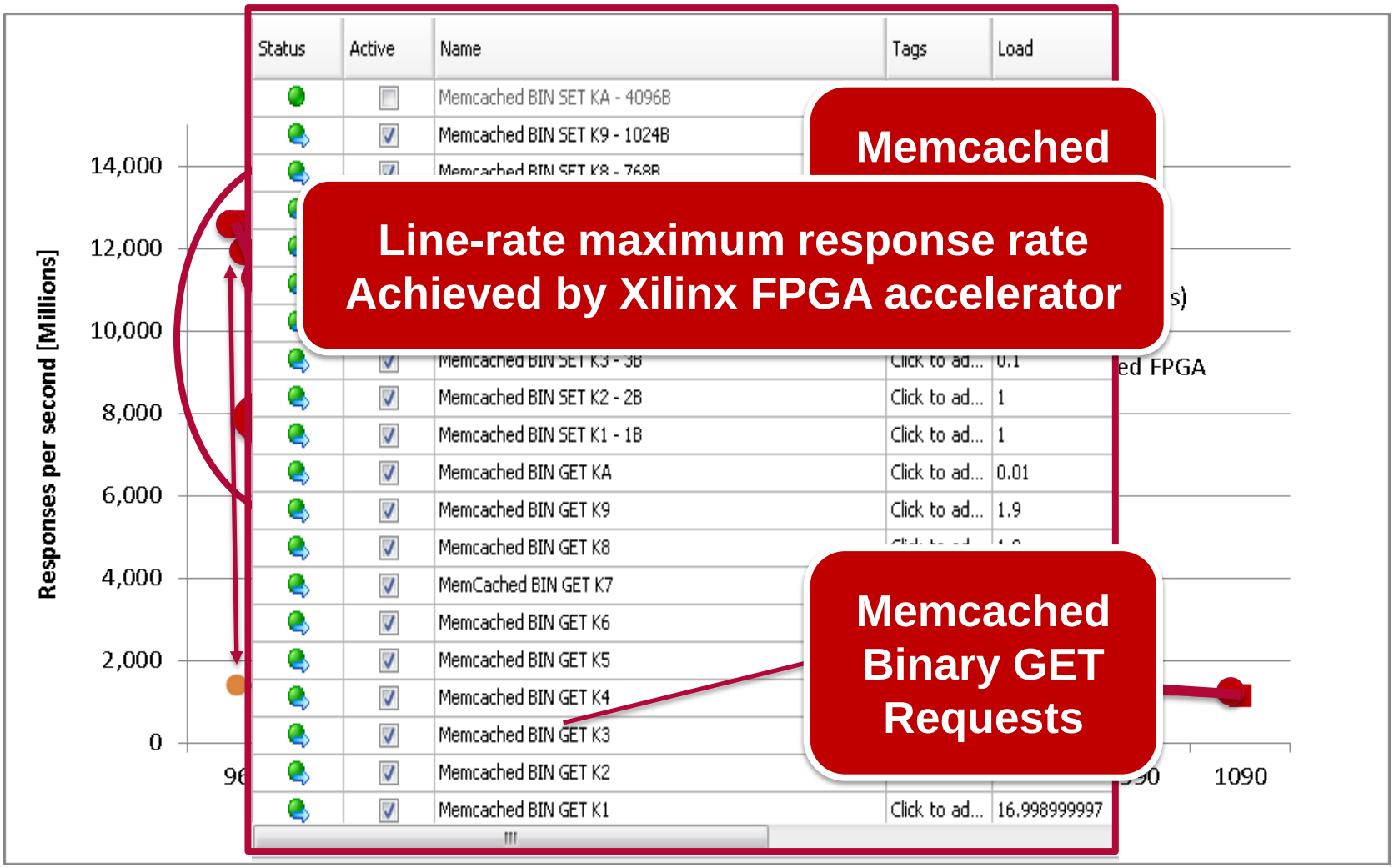
The Lab



Test Client Setup



Typical Facebook Request Distribution



Acknowledgements and References

- **Michaela Blott, Kimon Karris, Lisa Liu (Xilinx CTO Ireland)**
- **many students and interns**
- **Xilinx HLS team**

References:

- Hot Cloud 13**: Achieving 10Gbps Line-rate Key-value Stores with FPGAs,** Michaela Blott, Kimon Karras, Ling Liu, and Kees Vissers, *Xilinx Inc.;* Jeremia Bär and Zsolt István, *ETH Zürich*
- Hot Chips 2013: **Dataflow Architectures for 10Gbps Line-rate Key-value-Stores** Michaela Blott, Kees Vissers - Xilinx Research
- ISCA 2014: **A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services**, Andrew Putnam, Adrian M. Caulfield, Eric S. Chung, Derek Chiou, Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers,Gopi Prashanth Gopal, Jan Gray, Michael Haselman, Scott Hauck, Stephen Heil, Amir Hormati, Joo-Young Kim, Sitaram Lanka, James Larus, Eric Peterson, Simon Pope, Aaron Smith, Jason Thong, Phillip Yi, Xiao Doug Burger. *Microsoft*