OPENFABRICS ALLIANCE

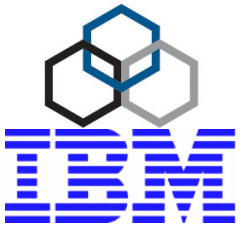
Energy efficient Large Layer 2 networks

Author: Ronald P. Luijten, DMA (Data Motion Architect) – IBM Research
Date: April 2011

DISCLAIMER:

This presentation is entirely Ronald's view and not necessarily that of IBM.

Rapidly changing technology and … Rapidly changing workloads



- Very soon, more data will be generated by devices than people
  - Smarter Planet, sensors
- DCN workloads already changing due to Twitter, Amazon, mobile devices
- µServer technology emerging now
  - ARM, PowerPC, X86
  - 4 core, 8 core available now
  - SOC designs: highly integrated
- New memory technologies emerging

Have you watched the µServer space lately?

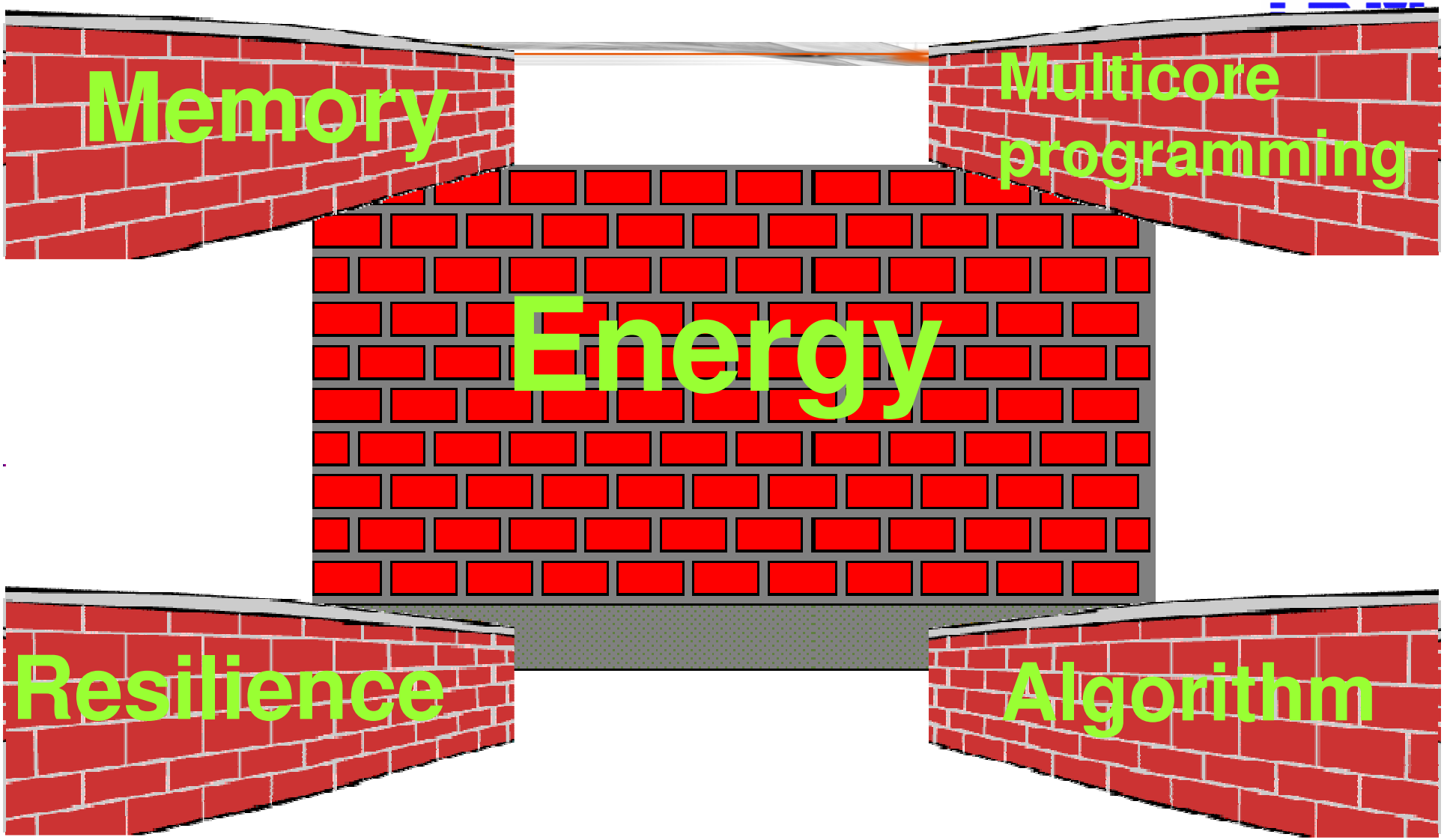
Incumbents dominate
Sustaining battles

Sustaining innovations

Disruptive innovations

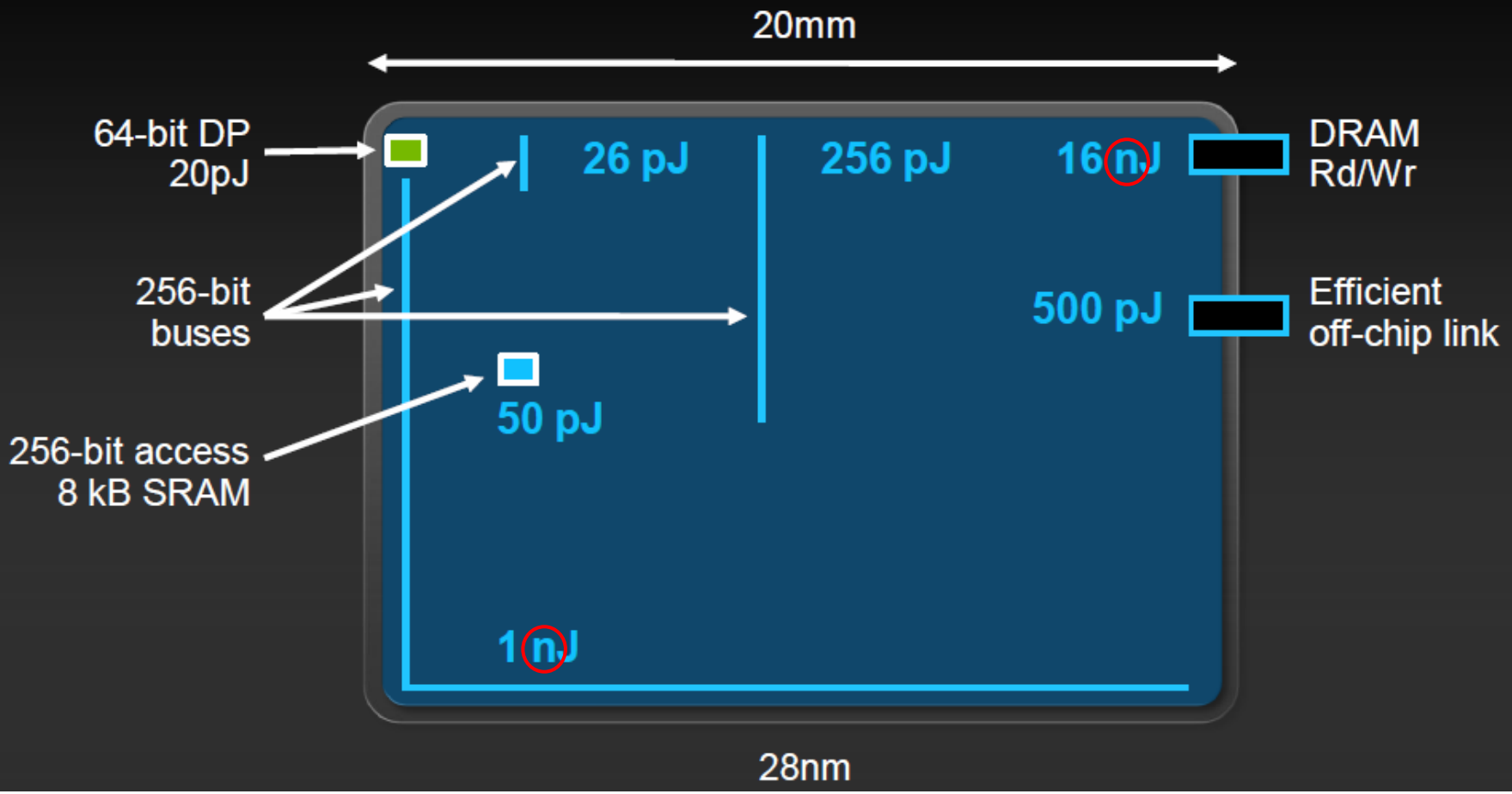
Entrants
typically win
at disruption

5 server walls



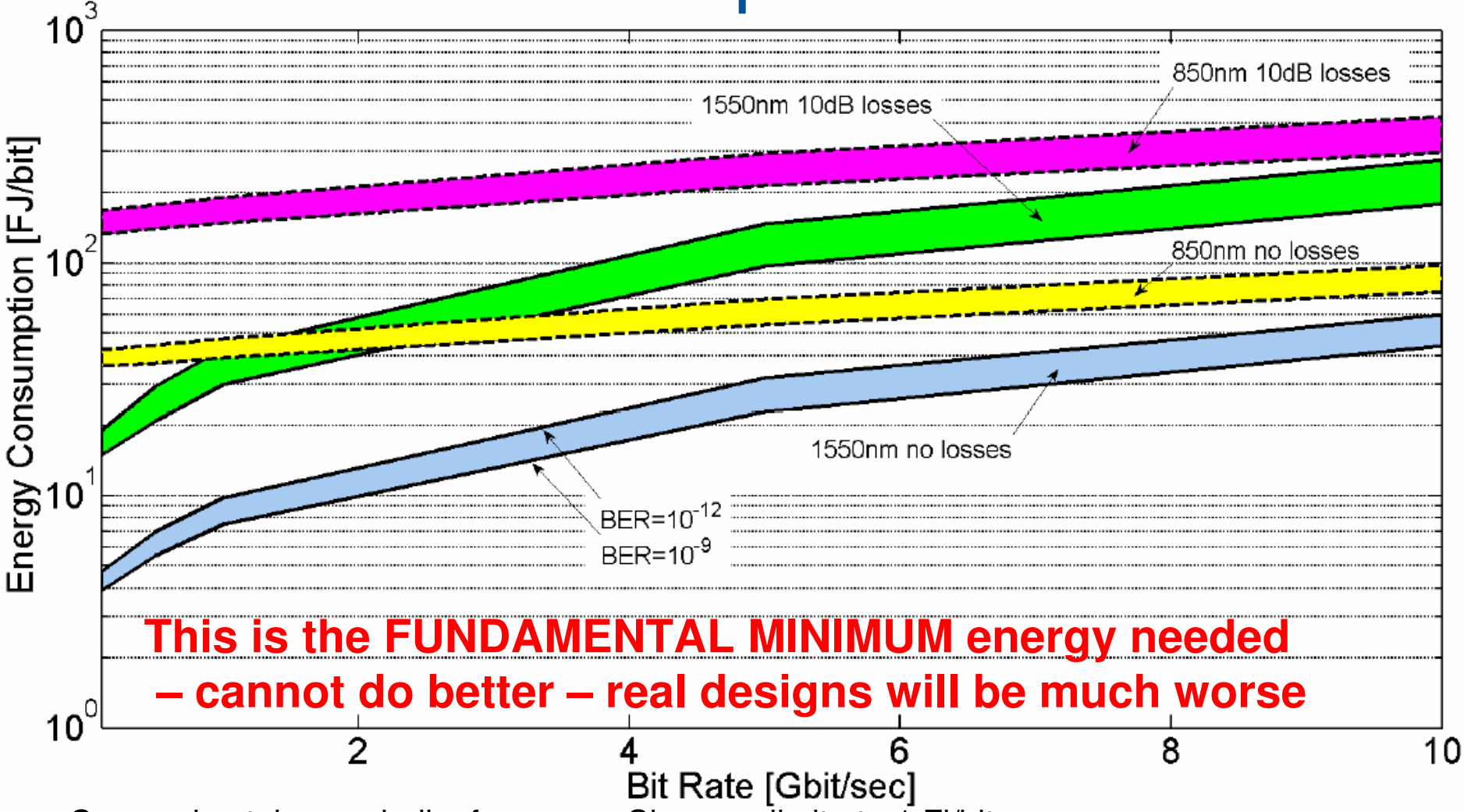
The High Cost of Data Movement

Fetching operands costs more than computing on them



Source: B. Dally, "GPU computing to Exascale and Beyond", SC10 Keynote, Nov 2011

CMOS 2 CMOS optical link



This is the FUNDAMENTAL MINIMUM energy needed – cannot do better – real designs will be much worse

Conversion takes up bulk of energy – Shannon limit at ~1 Fj/bit

Source: Prof. Harm Dorren, OECC2011

Observations - 1

- New class of workloads don't need SMPs
  - SMPs to stay for classical workloads – not a growth area
  - Programmers are learning how to scale non-SMP
  - Hadoop / MapReduce is only a first of new methods
    (right now seen as the hammer)
- BigData is coming – does not fit in cache
  - Putting ever more cores on a die not useful
  - New non-volatile memory types coming
- Strength of µServers is tight integration – gives significant power savings (fewer chip crossings)

Observations -2

- Energy+Memory wall makes µServers good choice
  (for the new workloads -balance of system)
- A compute core is rapidly becoming a commodity
- TCP/IP addresses and solves WAN issues
  - Long flight times, lossy, collaborative system
- DCN: lossless, very short flight times, single control domain: TCP/IP for DCN is overkill: waist of energy
- Cost of developing a new PHY is tremendous – need one common PHY for ENET, PCI, … for >100 Gbps generation

Implications (requirements) - 1

- Industry is well advanced on network convergence
  - DCB: Enet, FC, and… IB
- Savings in Capex, Opex (people and energy)

- **Next step**: converge PCI and DCB
  - More savings in Capex, Opex (Energy)
  - Intel Lightpeak is going there now

Implications (requirements) - 2

- The days of mostly large pkts will be over all too soon
- Does a sensor need 1500 Byte payload to send temperature?
- QED

Implications (requirements) - 3

- µServer trend causes explosion of endpoints
- Large DCnetworks needed
- Highly energy efficient
- TCP/IP is not
- Many topos, subnets - virtualized
- →Openflow, TRILL, mac-in-mac
- Lossless, QoS, Congestion Control, Adaptive Routing, transport layer
- Very tight, energy efficient integration needed

Implications (requirements) - 4

- Oh yes – Optical of course, for any distance

Questions?

papers (µServer; Photonics)

- **"Parallelism and Data Movement Characterization of contemporary Application Classes** ", Victoria Caparros Cabezas, Phillip Stanley-Marbell, to appear in ACM SPAA 2011, June 2011

- "**Quantitative Analysis of the Berkeley Dwarfs' Parallelism and Data Movement Properties**", Victoria Caparros Cabezas, Phillip Stanley-marbell, to appear in ACM CF 2011, May 2011

- "**Performance, Power, and Thermal Analysis of Low-Power Processors for Scale-Out Systems**", Phillip Stanley-Marbell, Victoria Caparros Cabezas, to appear in IEEE HPPAC 2011, May 2011

- "**Pinned to the Walls—Impact of Packaging and Application Properties on the Memory and Power Walls**", Phillip Stanley-Marbell, Victoria Caparros Cabezas, Ronald P. Luijten, submitted to IEEE ISLPED 2011 (undergoing review, needs acceptance before publication), Aug 2011.

"**Fundamental Bounds for Photonic Interconnects**", H. Dorren, P. Duan, O. Raz and R. Luijten, to appear in OECC2011, 4-8 July, Taiwan.