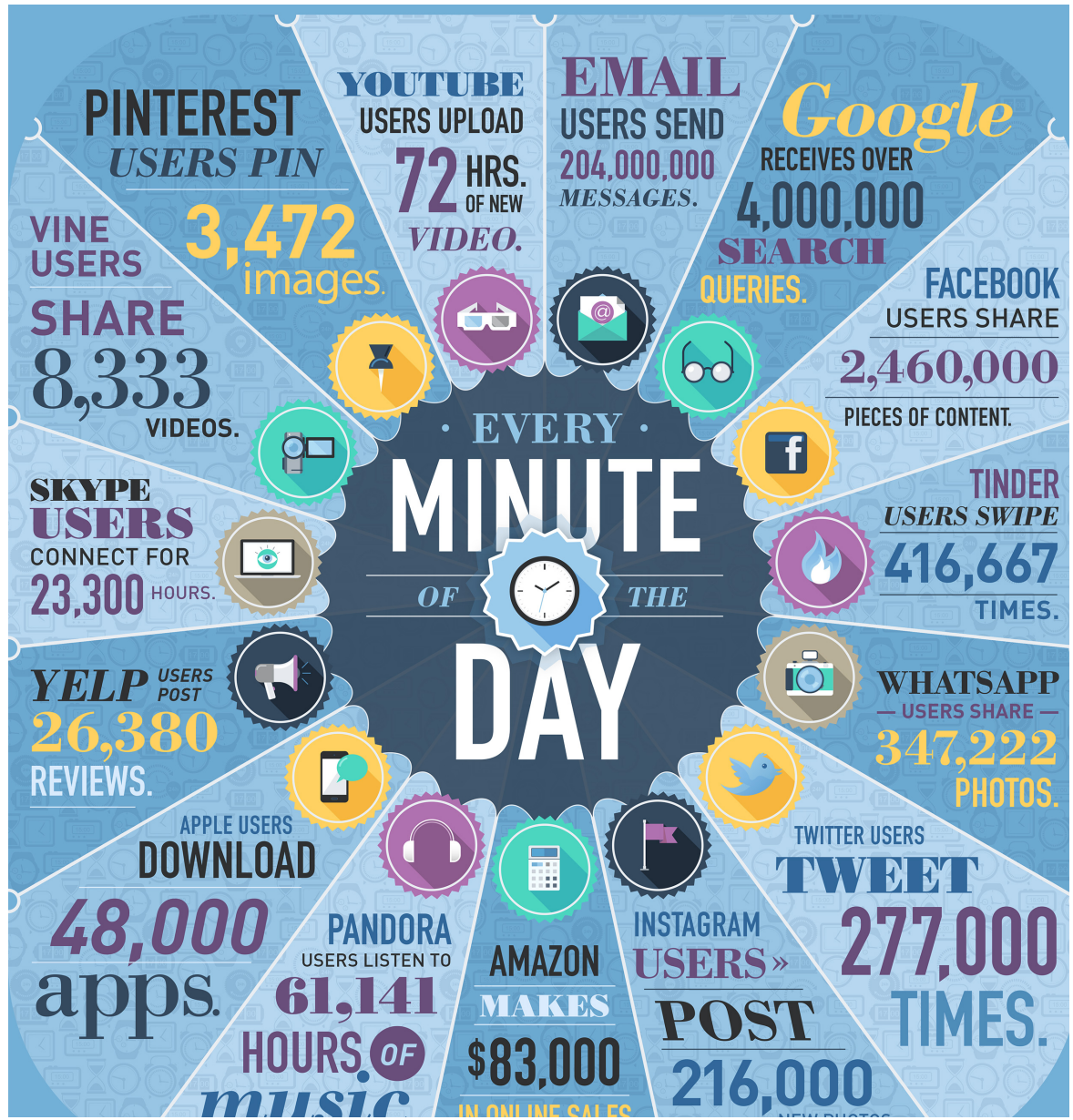
# Big Data
# Big Servers
# Big Trouble

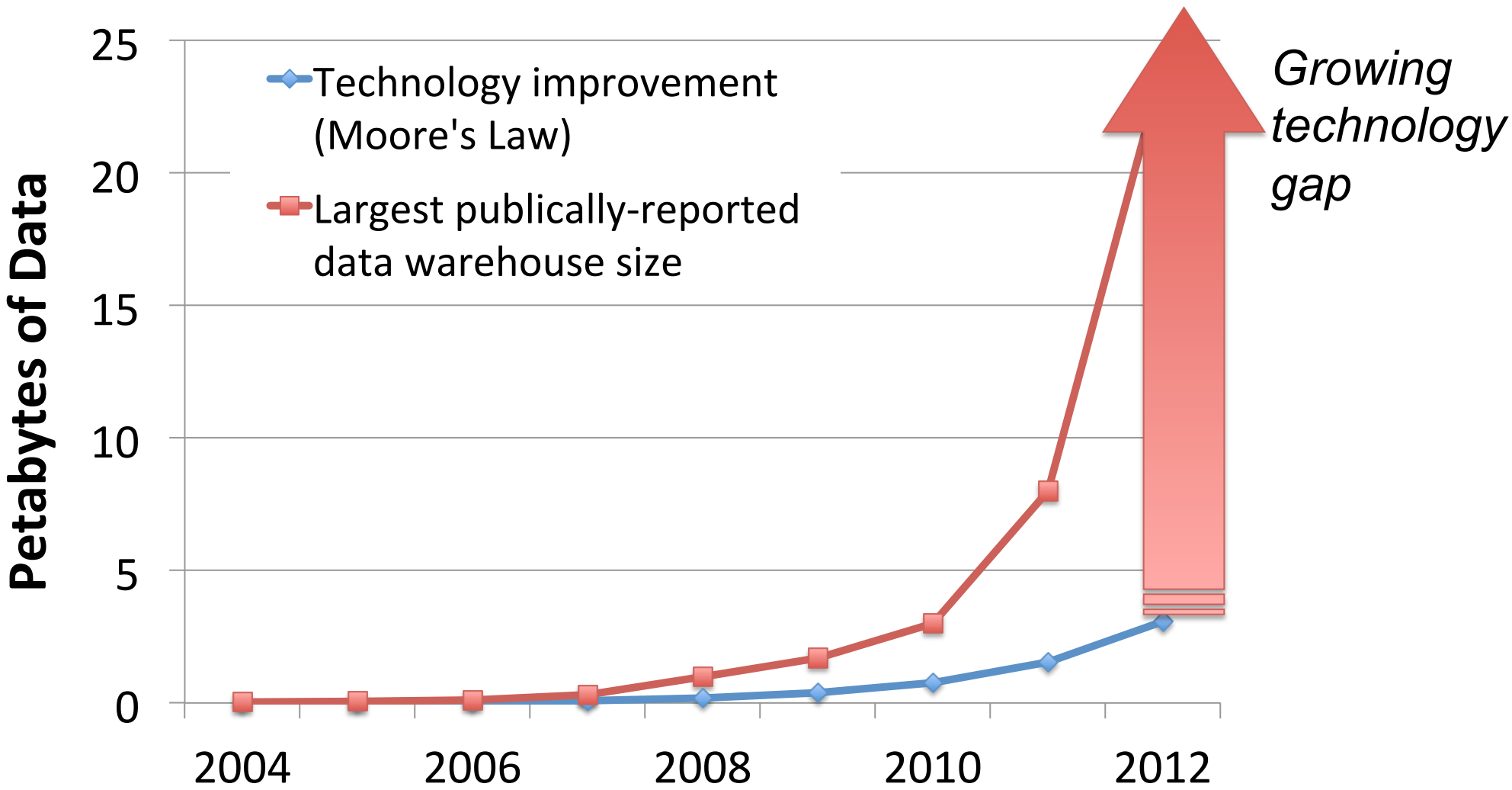
Boris Grot

University of Edinburgh

# The Big Data Explosion



# Data Grows, Technology Slows




WinterCorp Survey, www.wintercorp.com

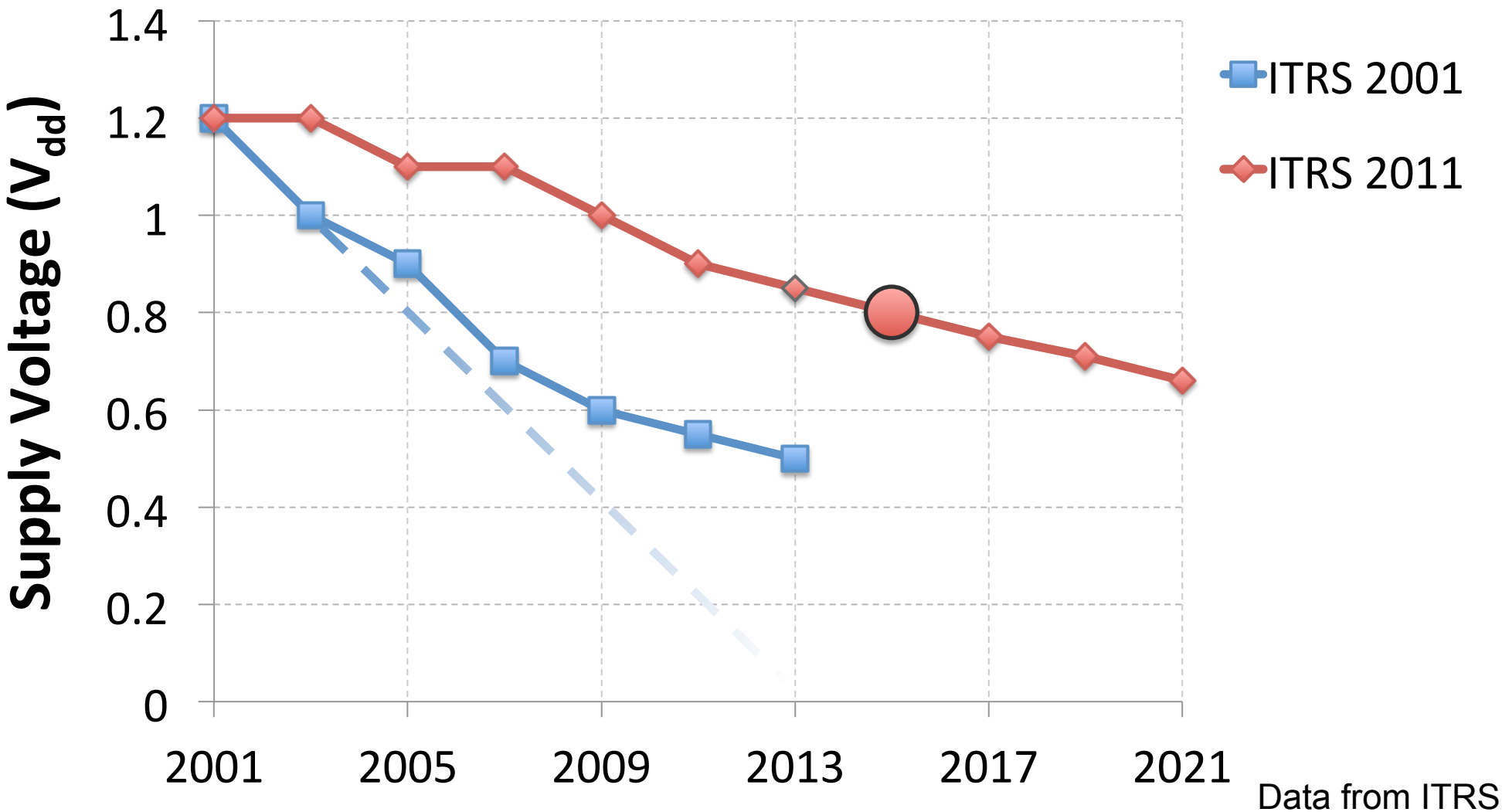
# Data-Centric IT Growing Fast

Source: James Hamilton, 2012


**Daily** IT growth in 2012 = IT first five years of business!

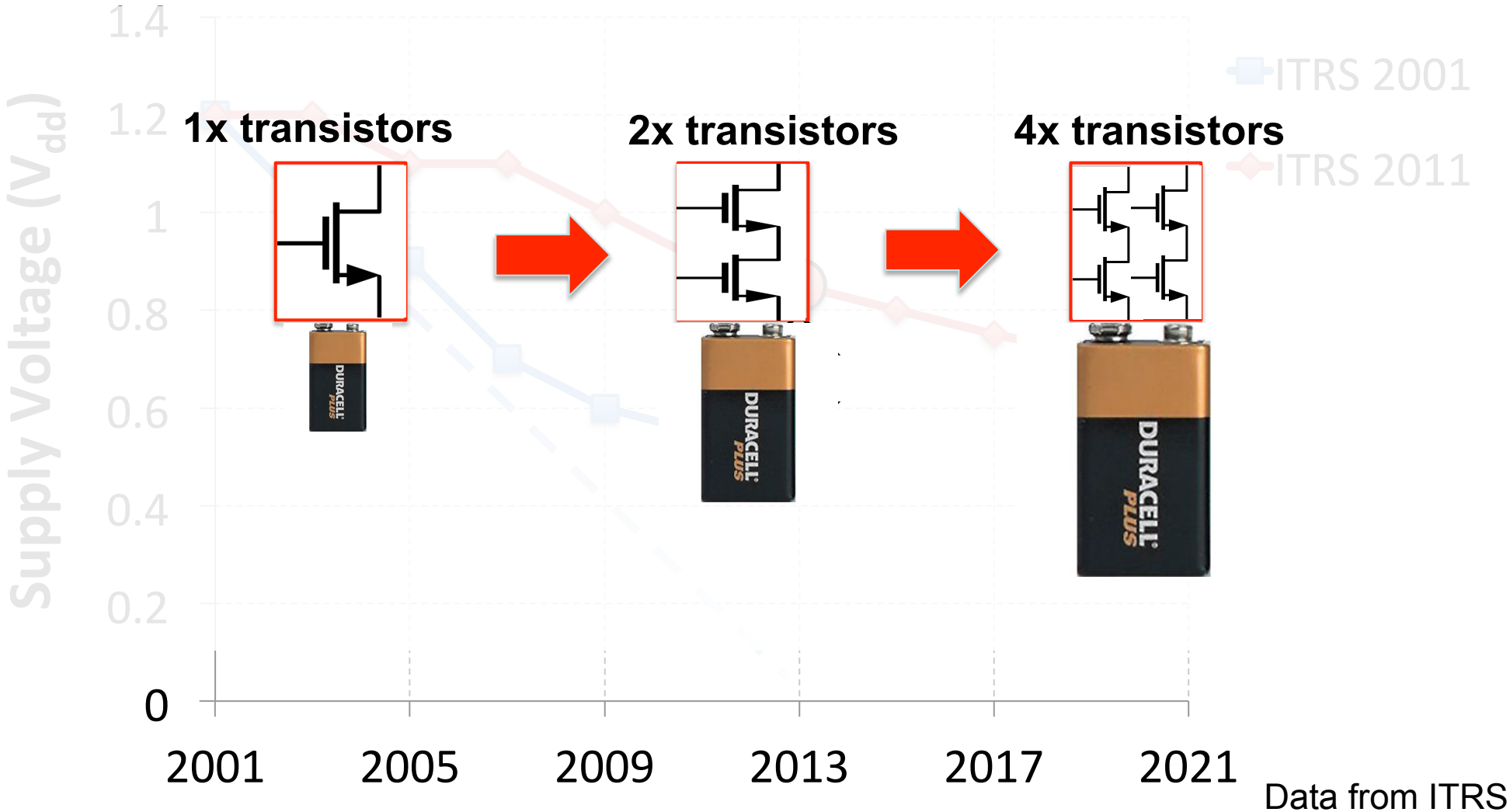
amazon web services

# Dennard Scaling is Dead




*Supply voltage scaling has slowed dramatically*

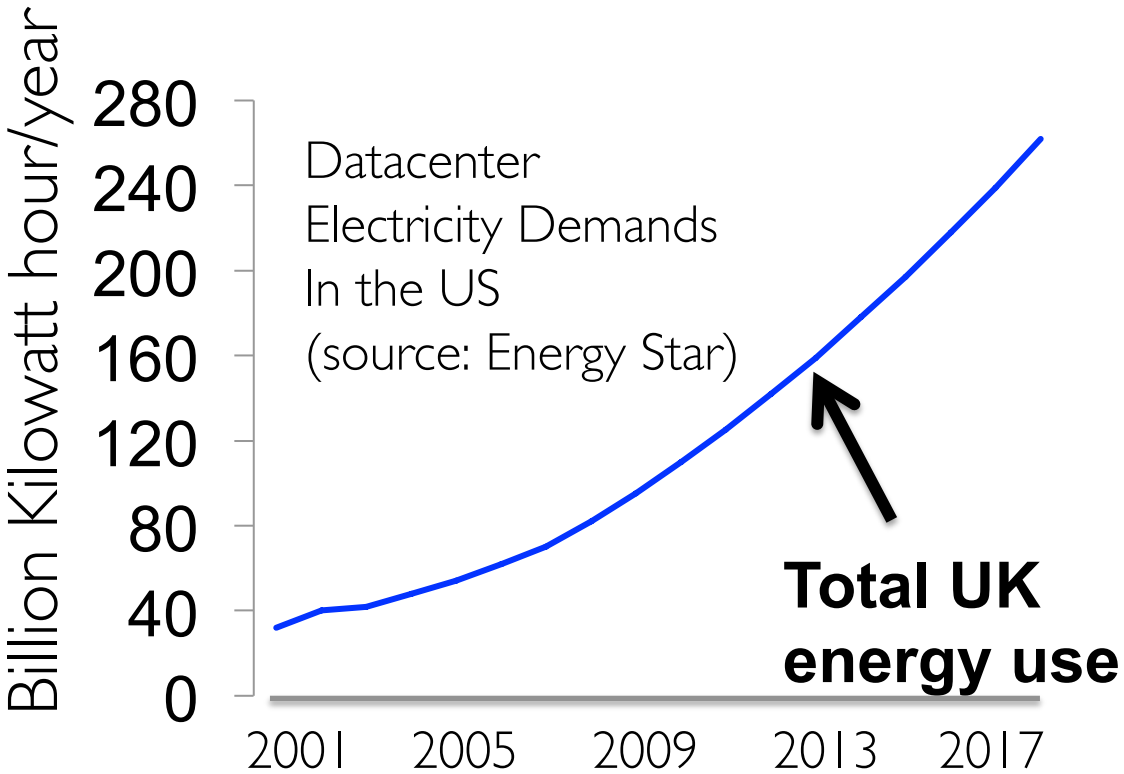
# Dennard Scaling is Dead

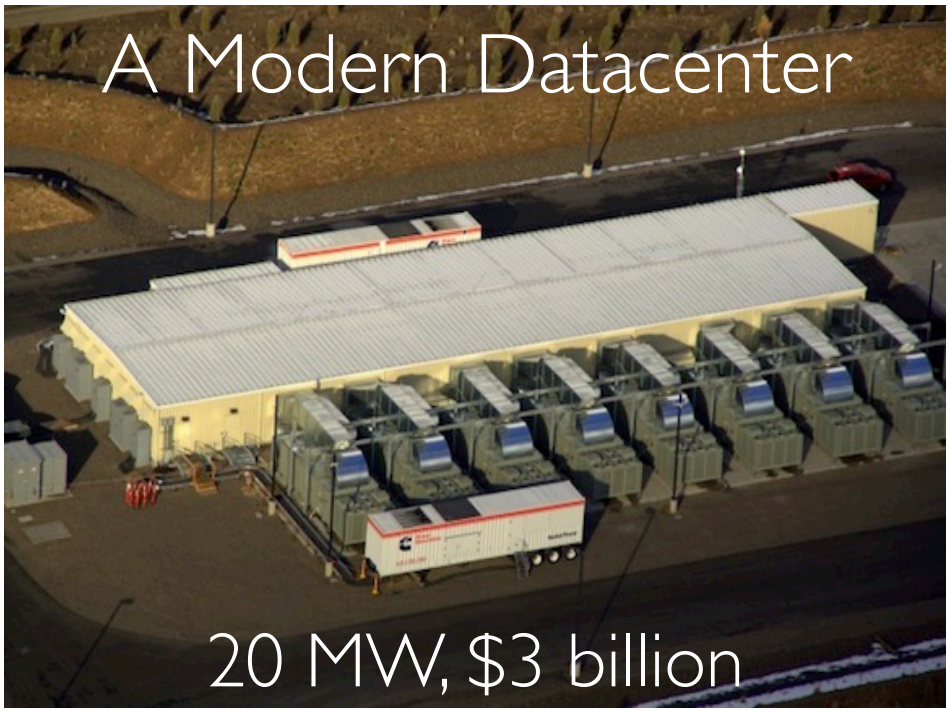



Data from ITRS

*Supply voltage scaling has slowed dramatically*

# Higher Demand + Lower Efficiency: Data-Centric IT Not Sustainable!




Today: datacenters draw 2% of global energy

Next decade: exponential growth if not mitigated

# NB: Datacenters are not Supercomputers

- Datacenters run data services at massive scale
- Fundamentally different design, reliability, performance, and energy efficiency targets
  - Datacenters optimize for TCO, not just performance
  - Datacenters embrace scale-out computing for cost and resilience


Supercomputer

≠

Datacenter

*Adapted from B. Falsafi, EcoCloud*

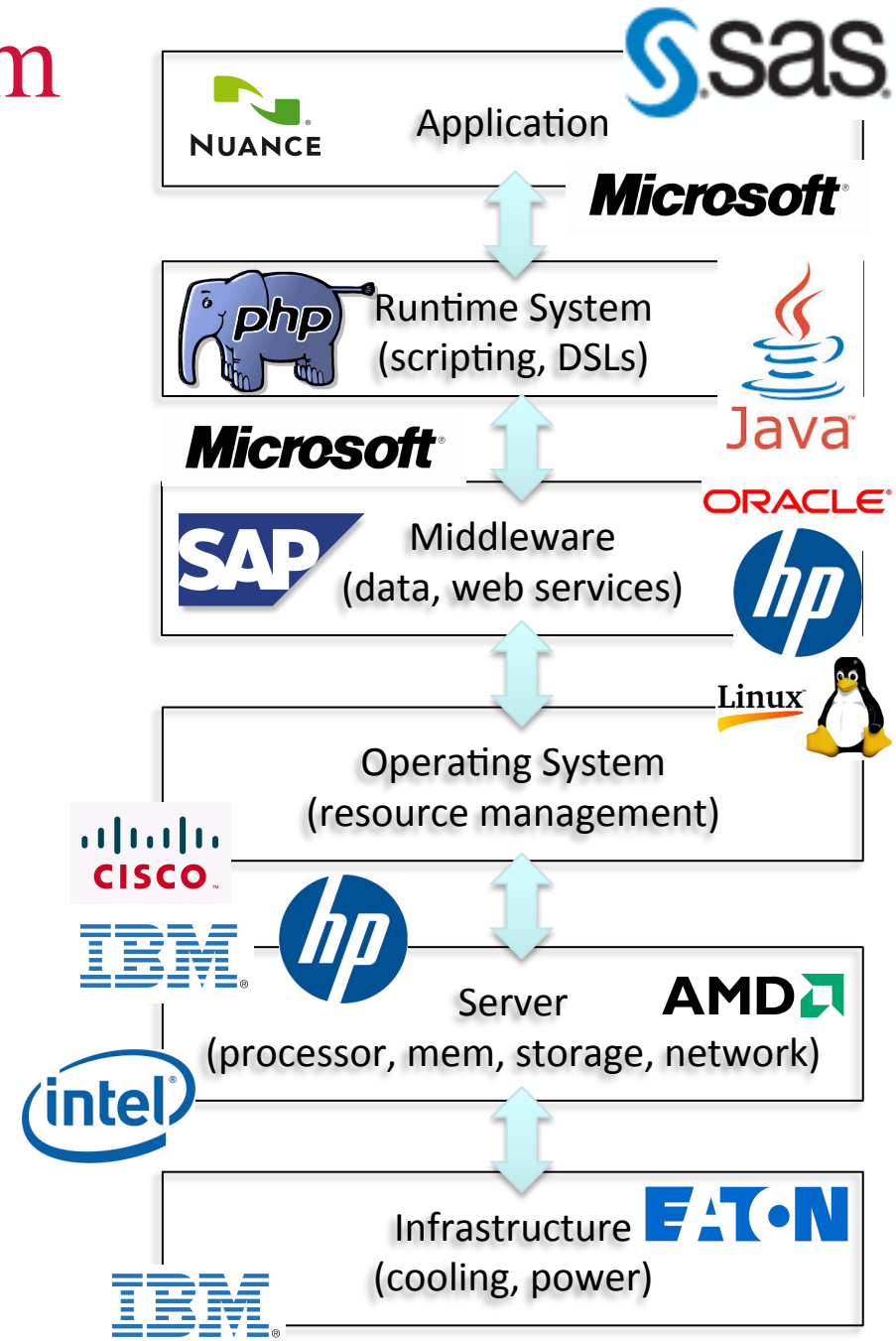
# Today's Server Ecosystem

## Conventional IT:

- Product based
- Strictly layered
- Near-neighbor optimization at best

## Datacenter operators + service providers (e.g., Amazon, Google):

- Can do cross-layer optimizations
- But,
  - Only limited to services of interest
  - Are limited in extent (e.g., software)
  - Monopolize (closed) technologies
  - Monopolize data




*Adapted from B. Falsafi, EcoCloud*

# The Way Forward

## Cross-layer optimization

- Performance predictability; widest range of power-perf modes
- E.g., rack-scale computing

## Specialization

- Fewer Joules/op; higher developer productivity
- E.g., accelerators, DSLs

## Integration

- TCO reduction, better latency, simpler management
- E.g., SoCs, memory fabrics

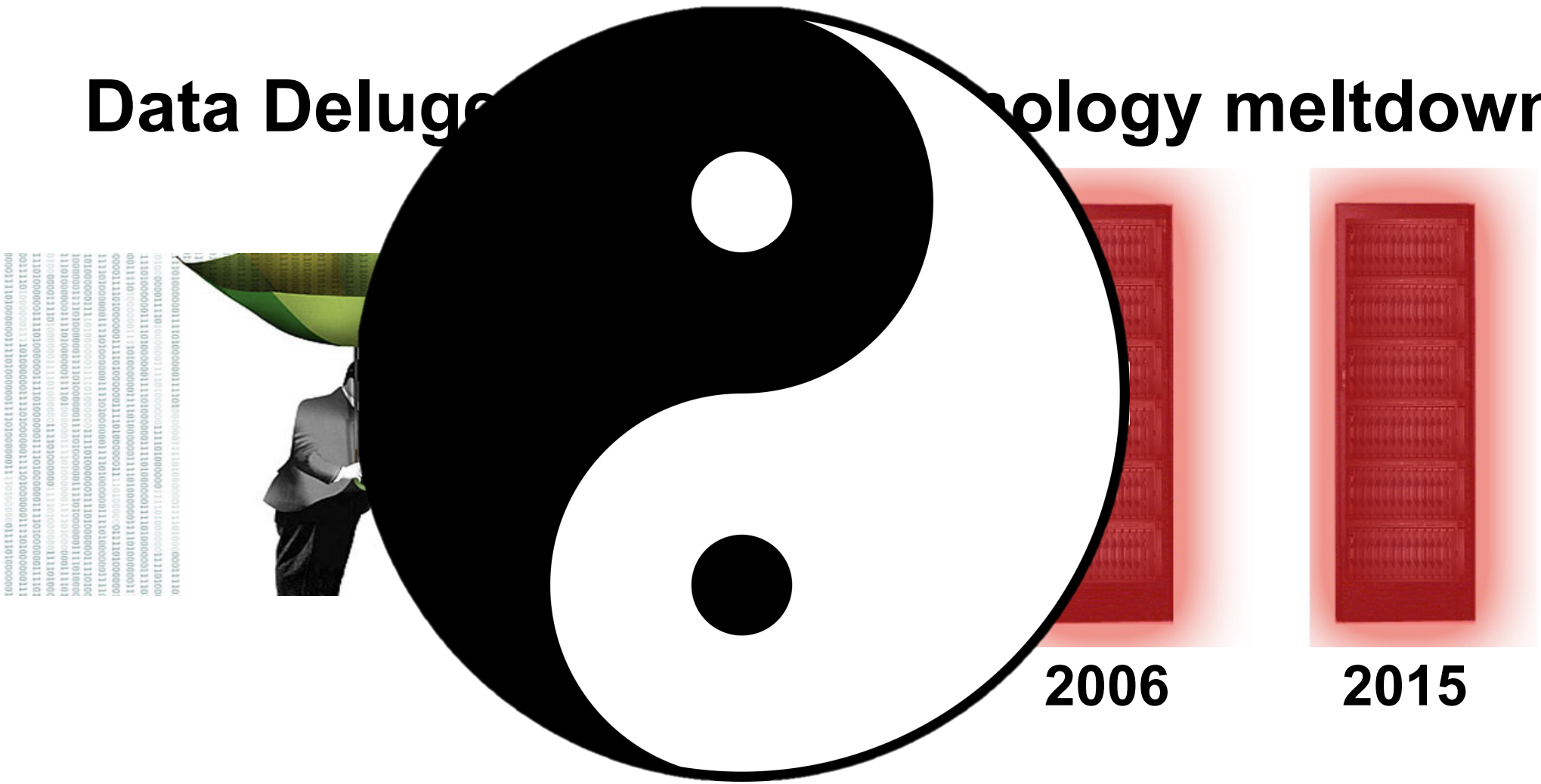
Full-stack Efficiency

Think

at

scale

# The Vision



# Scale-Out Processors

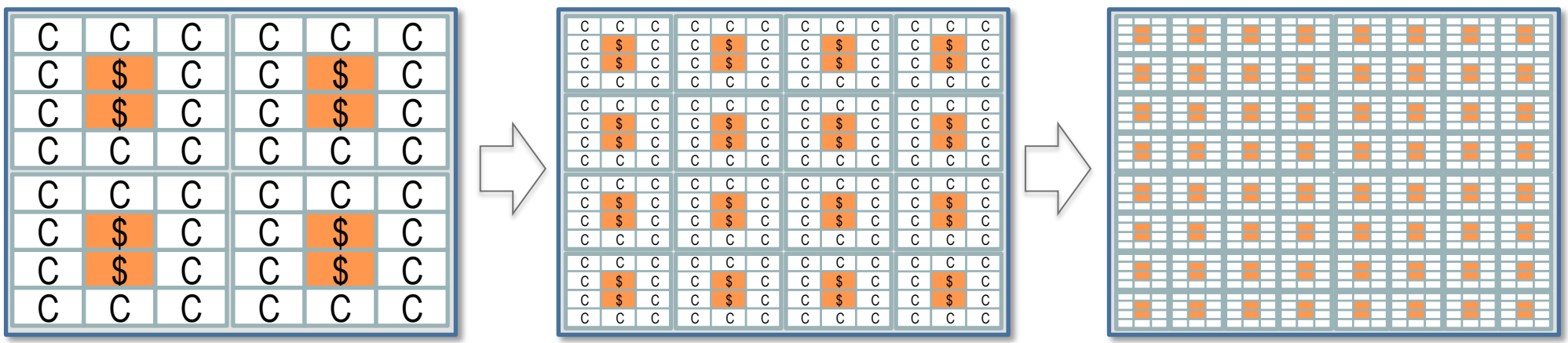
One or more pods

Each pod is a standalone server

- Runs a full software stack

No inter-pod connectivity or coherence

- Scalability and optimality across generations




Inherently optimal & scalable

# Memory: the New Efficiency Battleground

## DRAM:

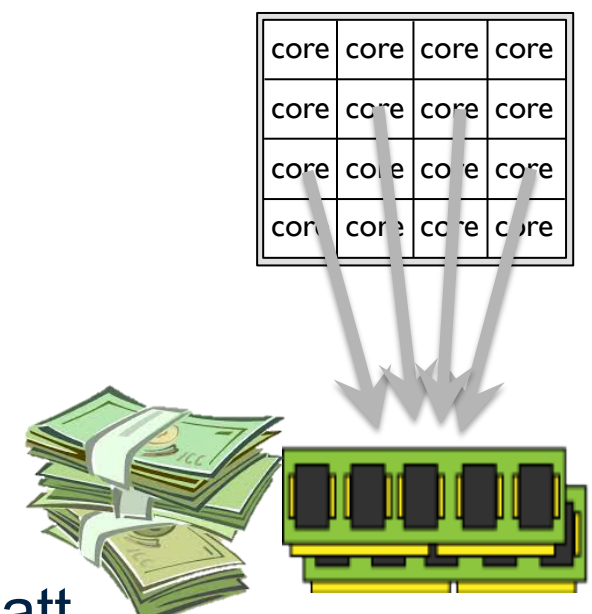
- Demand for capacity outpacing technology scaling
- Growing contributor to datacenter Total Cost of Ownership (TCO)


## Memory Desiderata

- High capacity, high BW, low power

## Existing Systems Fall Short

- DDR: at end of the road
- SerDes-based systems: poor capacity/Watt




*Must innovate in the memory system*

# Network: The Bottleneck for In-Memory Computing

Latency critical services

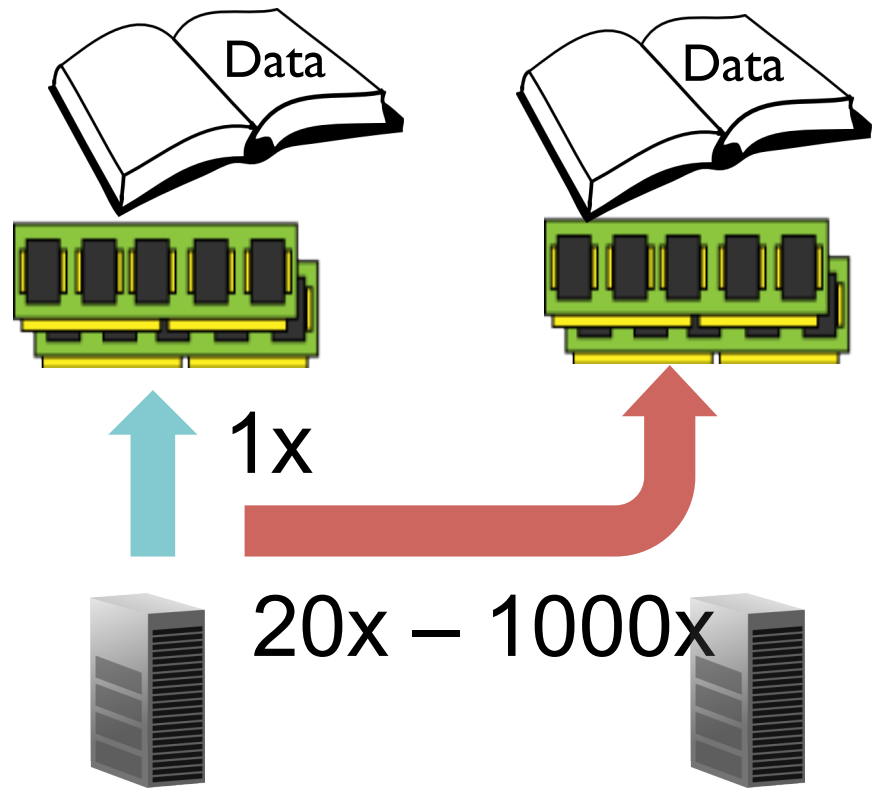
- Graphs, KV, DB: disk→DRAM

Vast datasets → distribute

- Often within rack

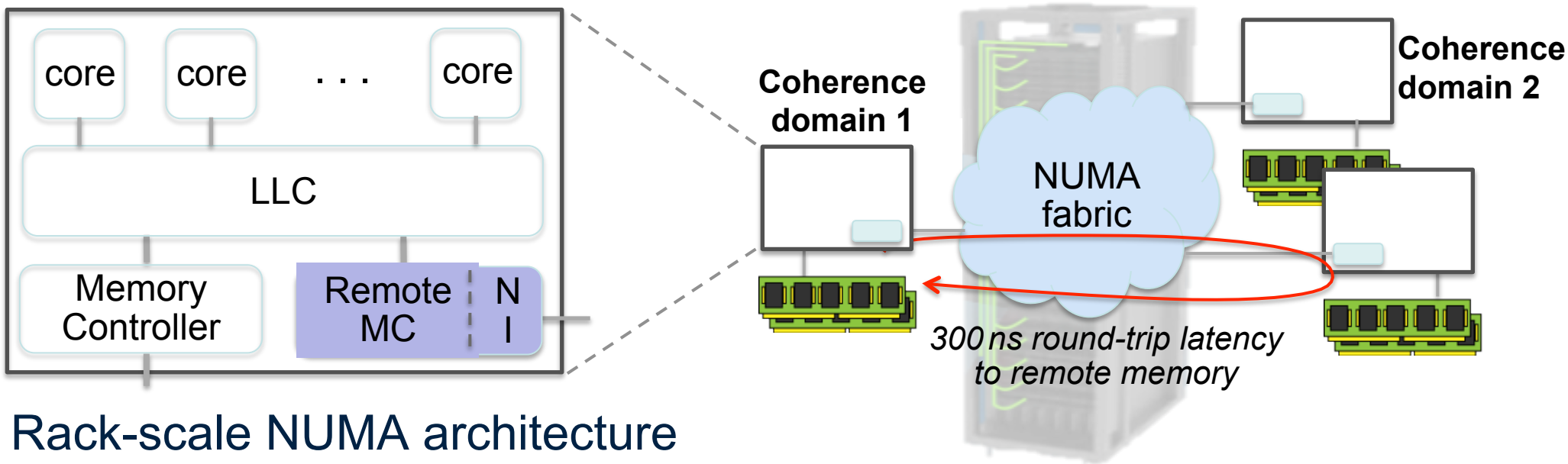
Today's networks:

✘ Latency 20x-1000x of local DRAM




Remote access latency >> local access latency

# Scale-Out NUMA (soNUMA)




## Rack-scale NUMA architecture

- Memory fabric of SoC servers
- Global virtual address space (NUMA, not cache-coherent)

## RDMA-inspired programming model

- HW-supported remote read, write, and atomics. Rest in SW

## Remote Memory Controller (RMC)

**soNUMA addresses all sources of latency**

# Thank you!

*Cross-layer optimization*

*Specialization*

*Integration*

Think

at

scale

# Questions?

***inf.ed.ac.uk/bgrot***