Cutting the energy costs of TV by a factor of five,by understanding the viewing figures for the top ten.

Jon Crowcroft, Andrew Moore, Nishanth Sastry (Kings, London) & Gianfranco Nencioni (Uni Pisa)&Jigna Chandaria (BBC) & The INTERNET (Intelligent Energy Aware Networking) Project

Cambridge University Computer Laboratory

INTERNET Project Background

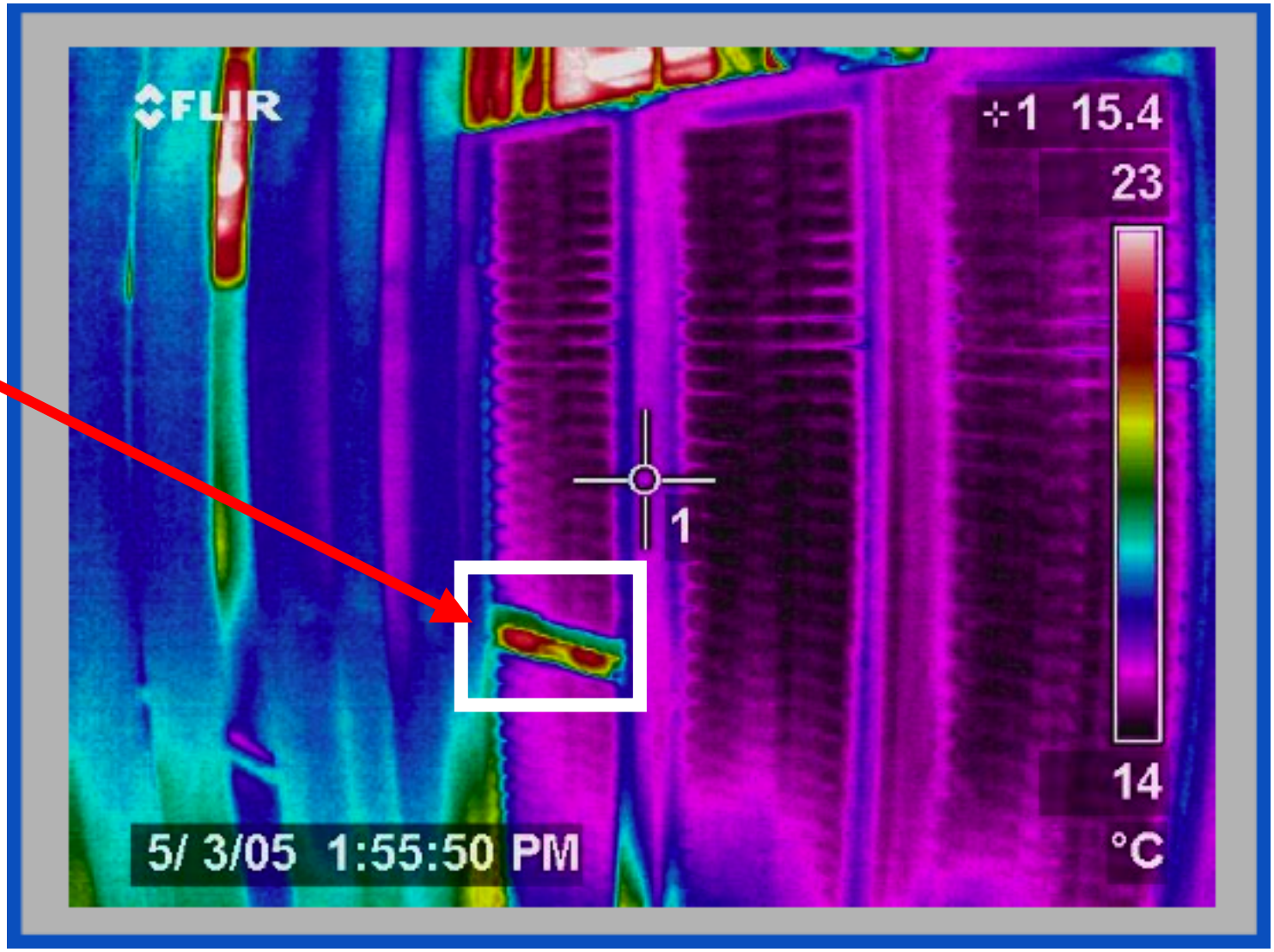
- 5 year project with industry including
  - Providers - e.g. BT
  - Users - e.g. BBC
  - Vendors - e.g. Cisco
- Look at reducing Carbon Footprint of Net
  - Goal - 10 fold reduction
  - Much through hardware, but also
  - Smart optimisation…

General Work Areas

1. Switches/Switchlets/Control Planes
2. Data Center Migration
    1. routing&addressing protocol implications
    2. Multipath transport
3. Optimising TV Distribution Energy Costs

Thermal Image of Typical Data Centre Rack

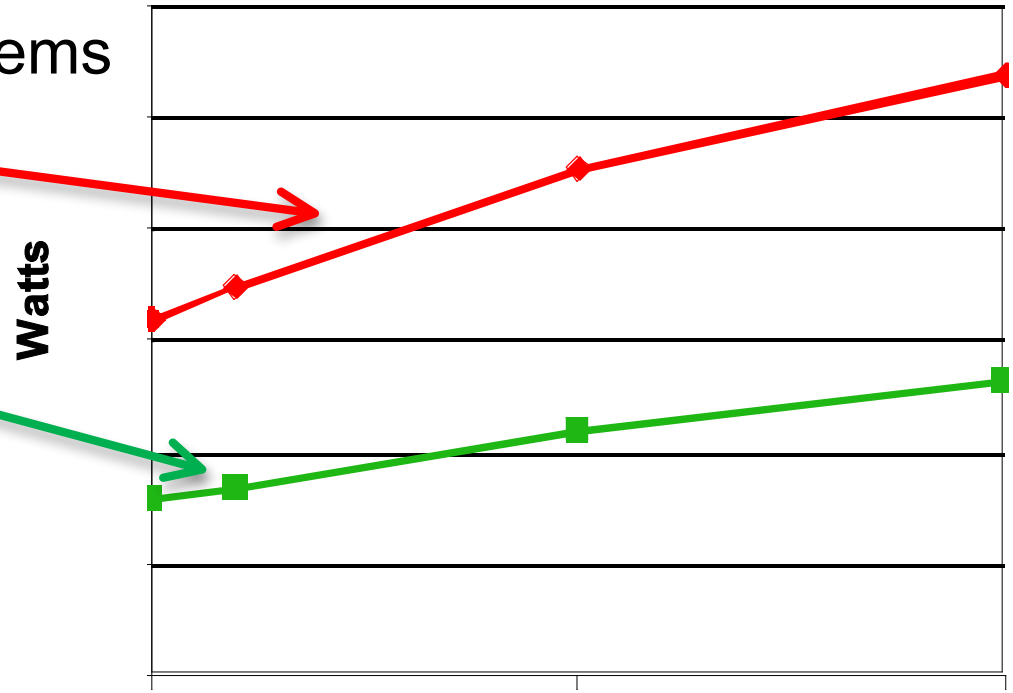
Rack Switch



M. K. Patterson, A. Pratt, P. Kumar,
"From UPS to Silicon: an end-to-end evaluation of datacenter efficiency", Intel Corporation

Motivating Consolidation

- SPECpower: two best systems
  - Two 3.0-GHz Xeons, 16 GB DRAM, 1 Disk
  - One 2.4-GHz Xeon, 8 GB DRAM, 1 Disk
- 50% utilization ➔ 85% Peak Power
- 10% ➔ 65% Peak Power
- Save 75% power if consolidate & turn off

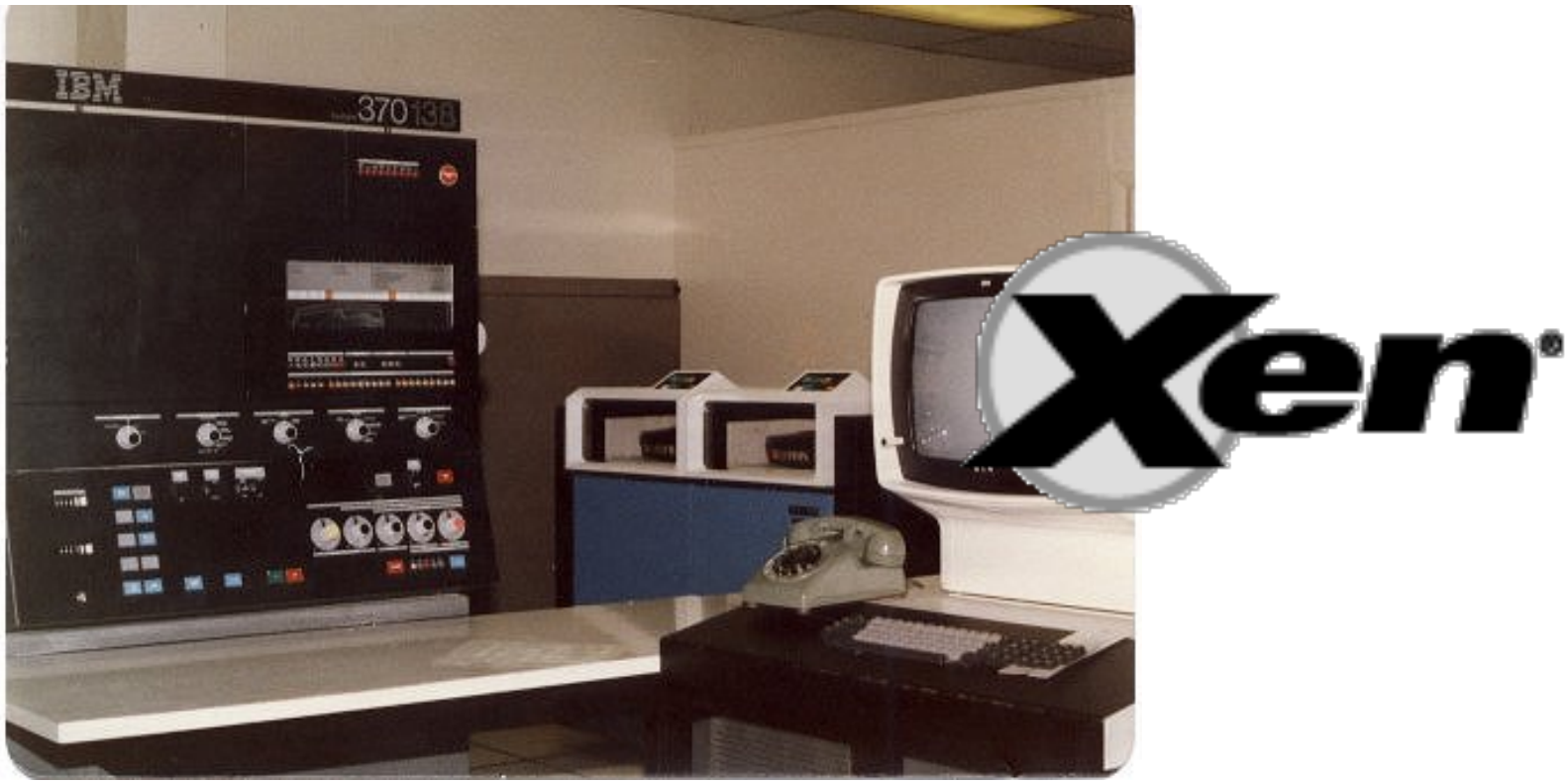


1 computer @ 50% = 225 W

vs 5 computers @ 10% = 870 W

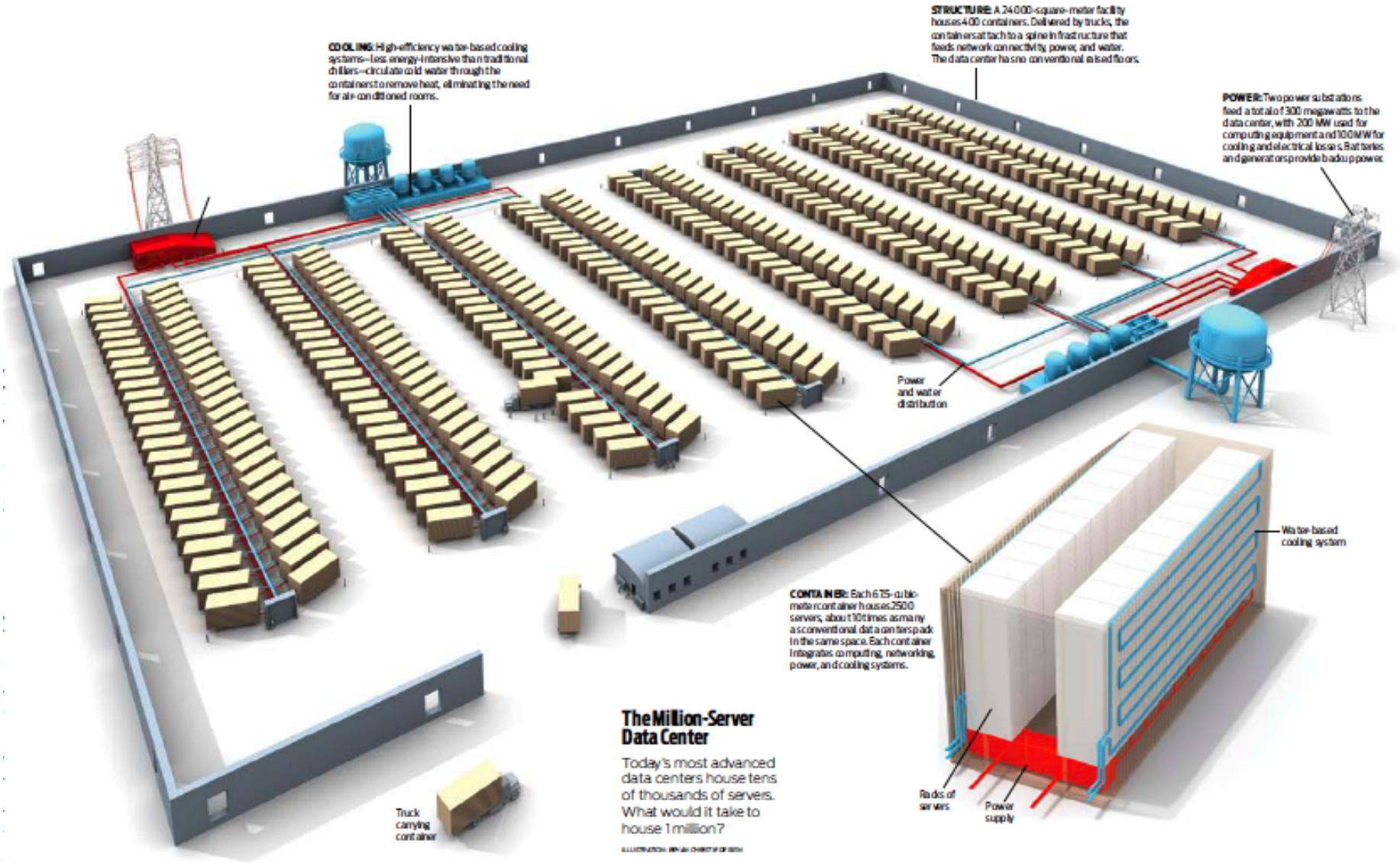
Better to have one computer at 50% utilization than five computers at 10% utilization: Save £ via consolidation (Saving £s on machines *and* power)

Lets *consolidate* like its 1969

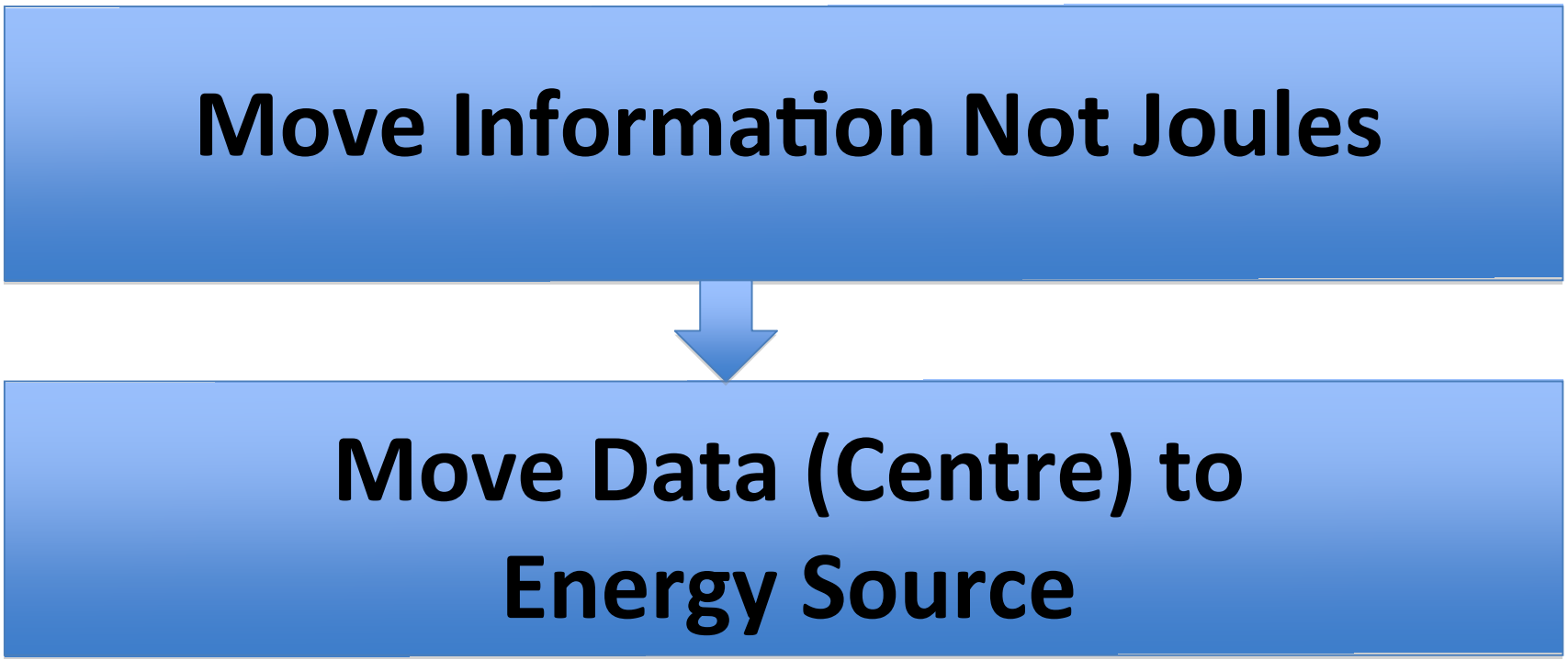


But *saving server power* is not the only fruit…

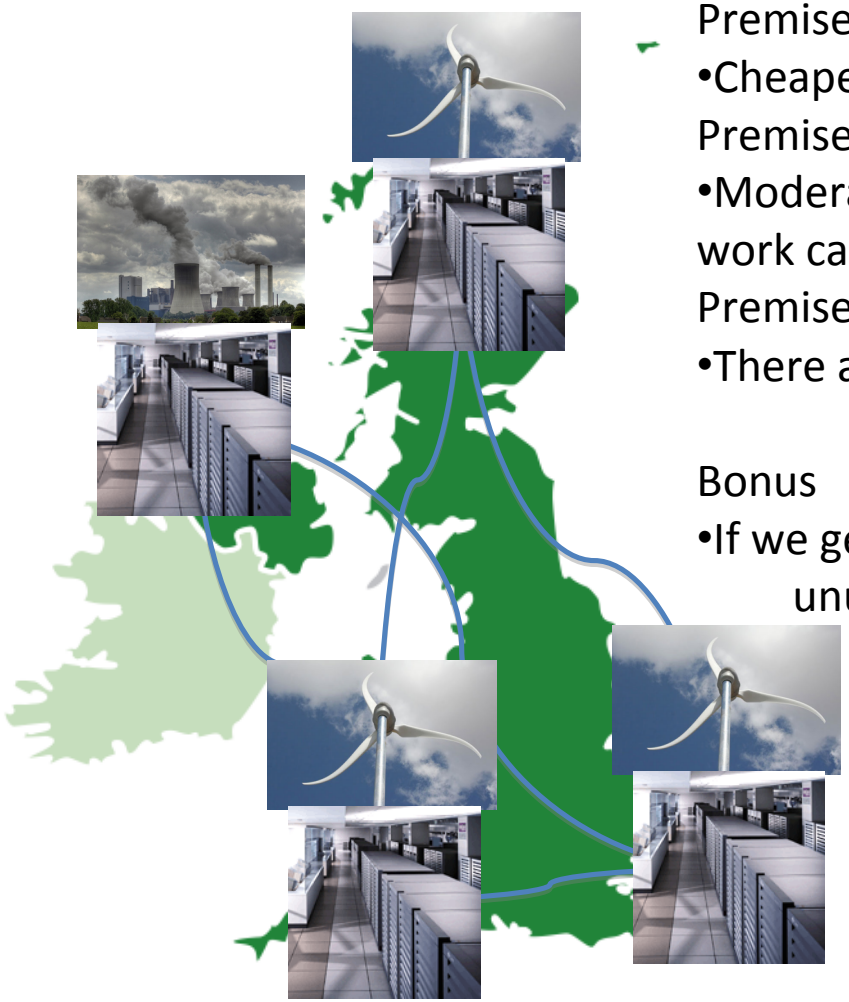
Microsoft's Chicago Modular Datacenter



INTelligent Energy awaRe NETworks



“Supply-following” Data Centers



Premise 1

•Cheaper to lay and maintain fiber than powerlines

Premise 2

•Moderate/Sufficient diversity in energy sources means, if work can follow supply, work can be continuous

Premise 3

•There are lots of places to make an energy improvement

Bonus

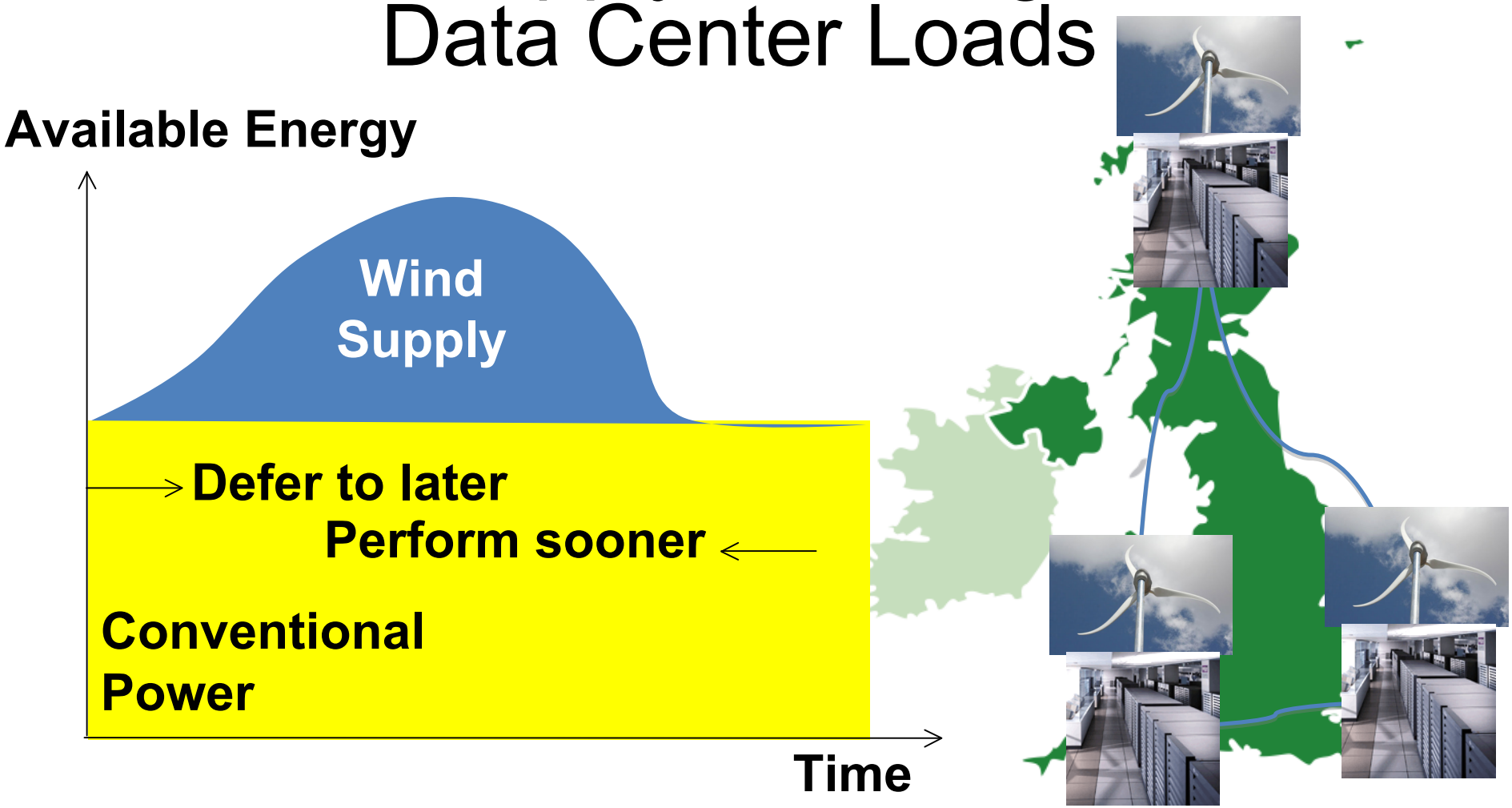
•If we get this right; can we run a data center on unused sustainable energy?

…Free-lunch Computing…

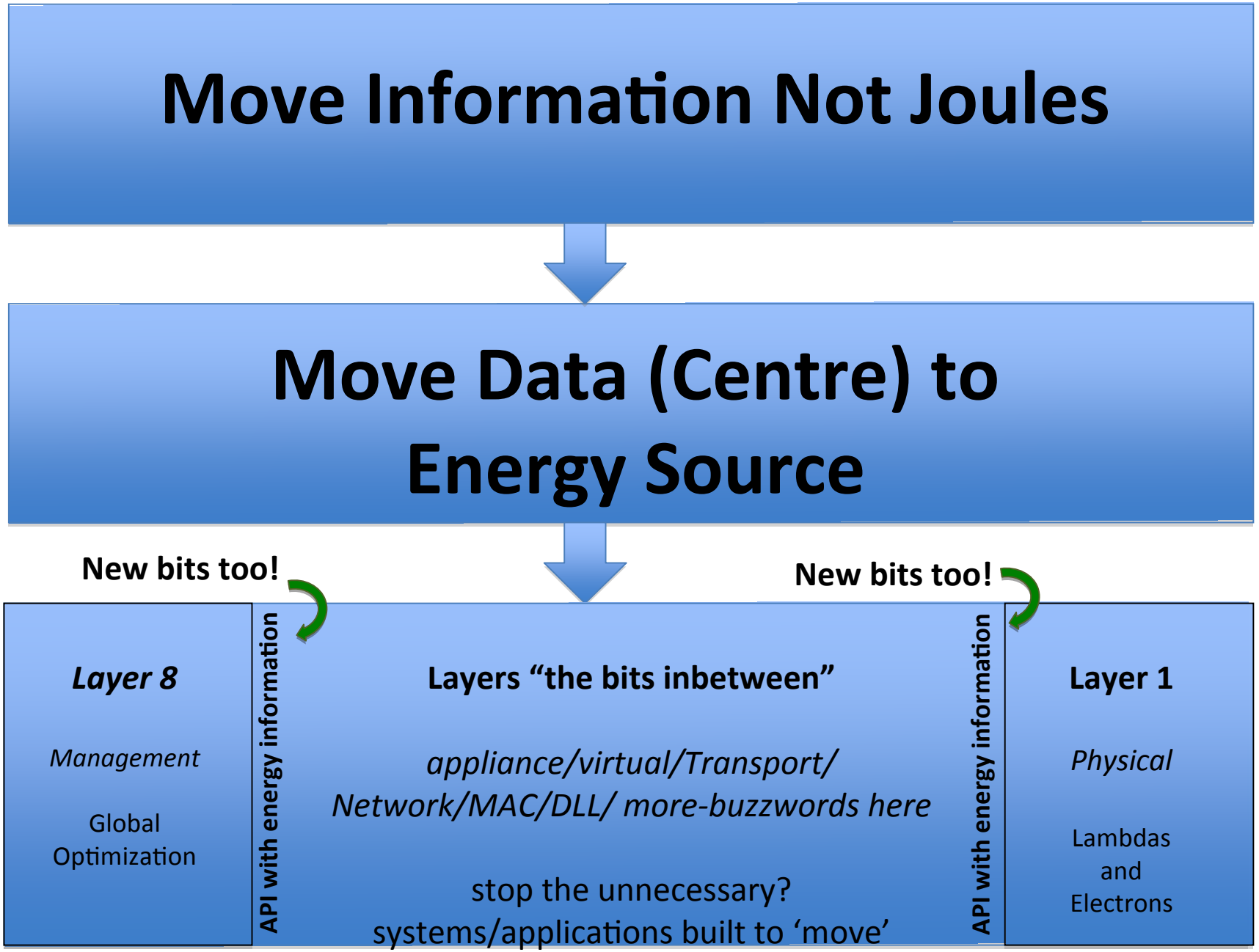
Andy Hopper

A logical conclusion leads a micro datacenter in every wind turbine.

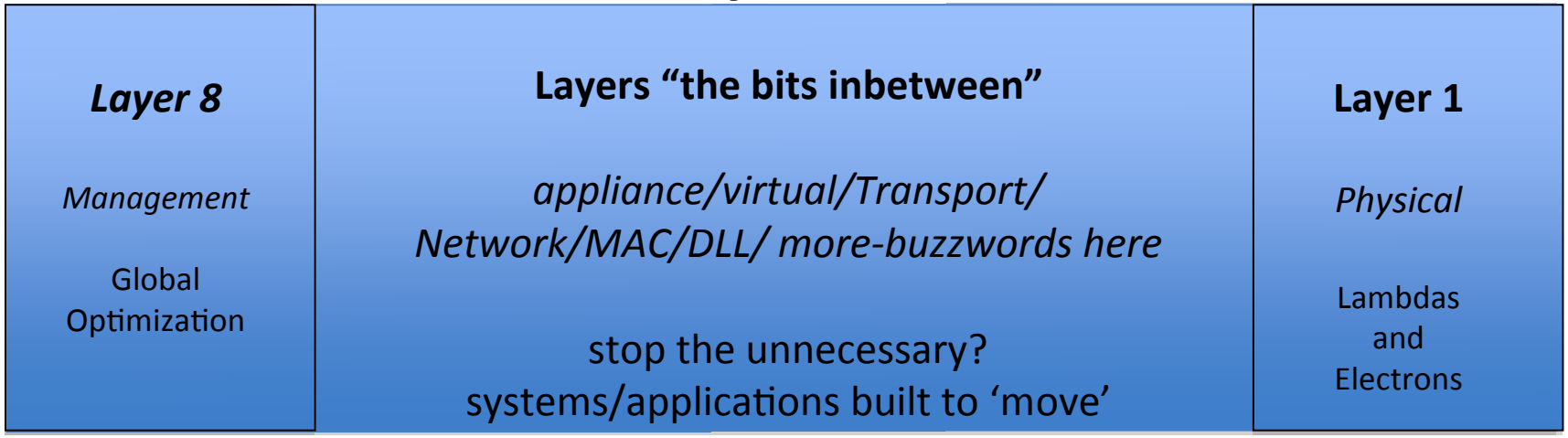
“Supply-following” Data Center Loads



- “Make hay while the Sun shines”: Do more when supply is available, defer when it is not
- Workload awareness is essential
- Better Forecasting means Better effectiveness


Move Information Not Joules
Move Data (Centre) to Energy Source
New bits too!
New bits too!
Layer 8
Management
Global Optimization
API with energy information
Layers "the bits inbetween"
appliance/virtual/Transport/ Network/MAC/DLL/ more-buzzwords here
stop the unnecessary?
systems/applications built to 'move'
API with energy information
Layer 1
Physical
Lambdas and Electrons

Layers…



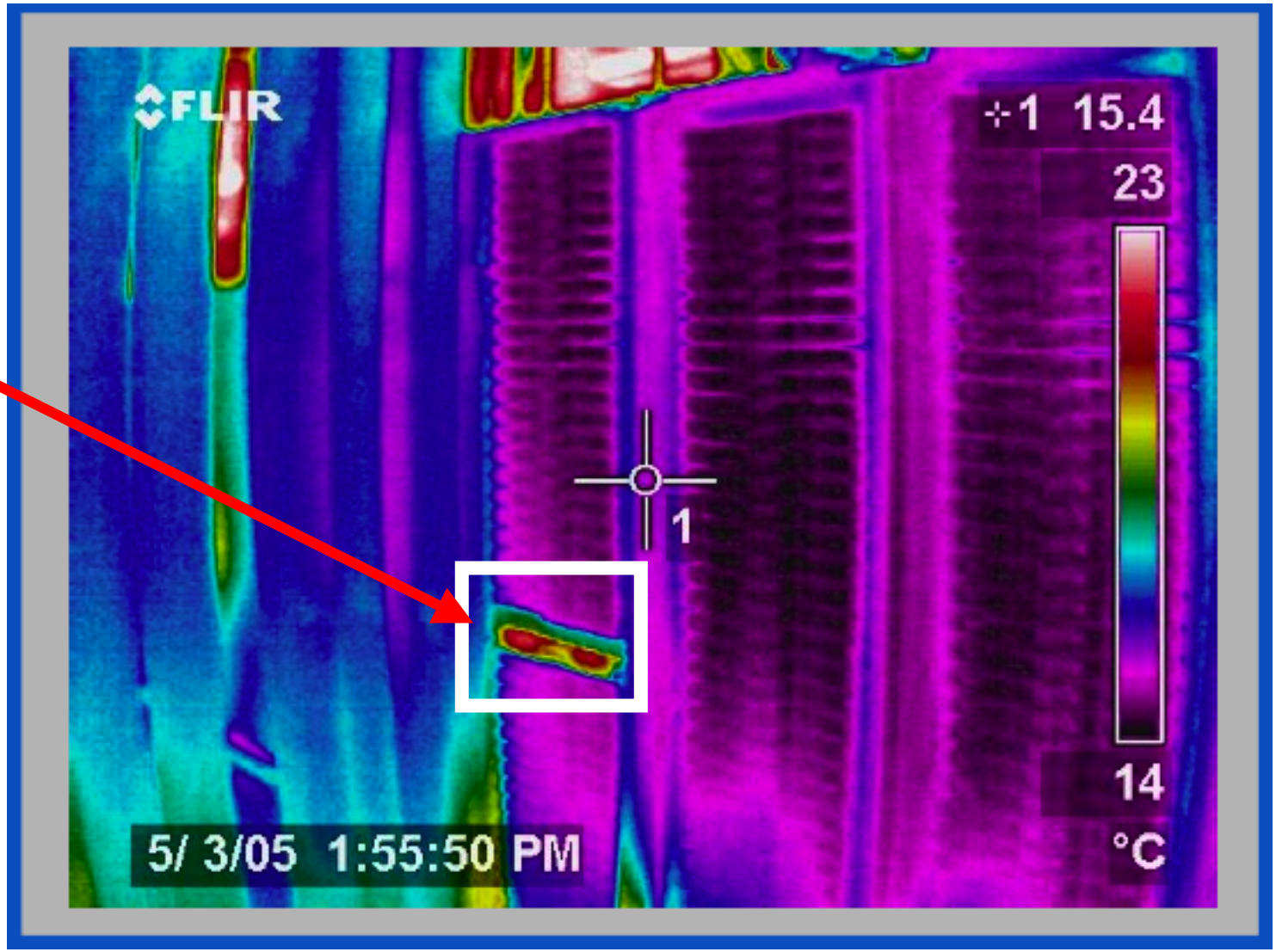
Application: Service Level Agreements for Migration

Transport: Multipath TCP to minimize impact on users

Network: Why should we build datacenters like mini-Internets anyway?

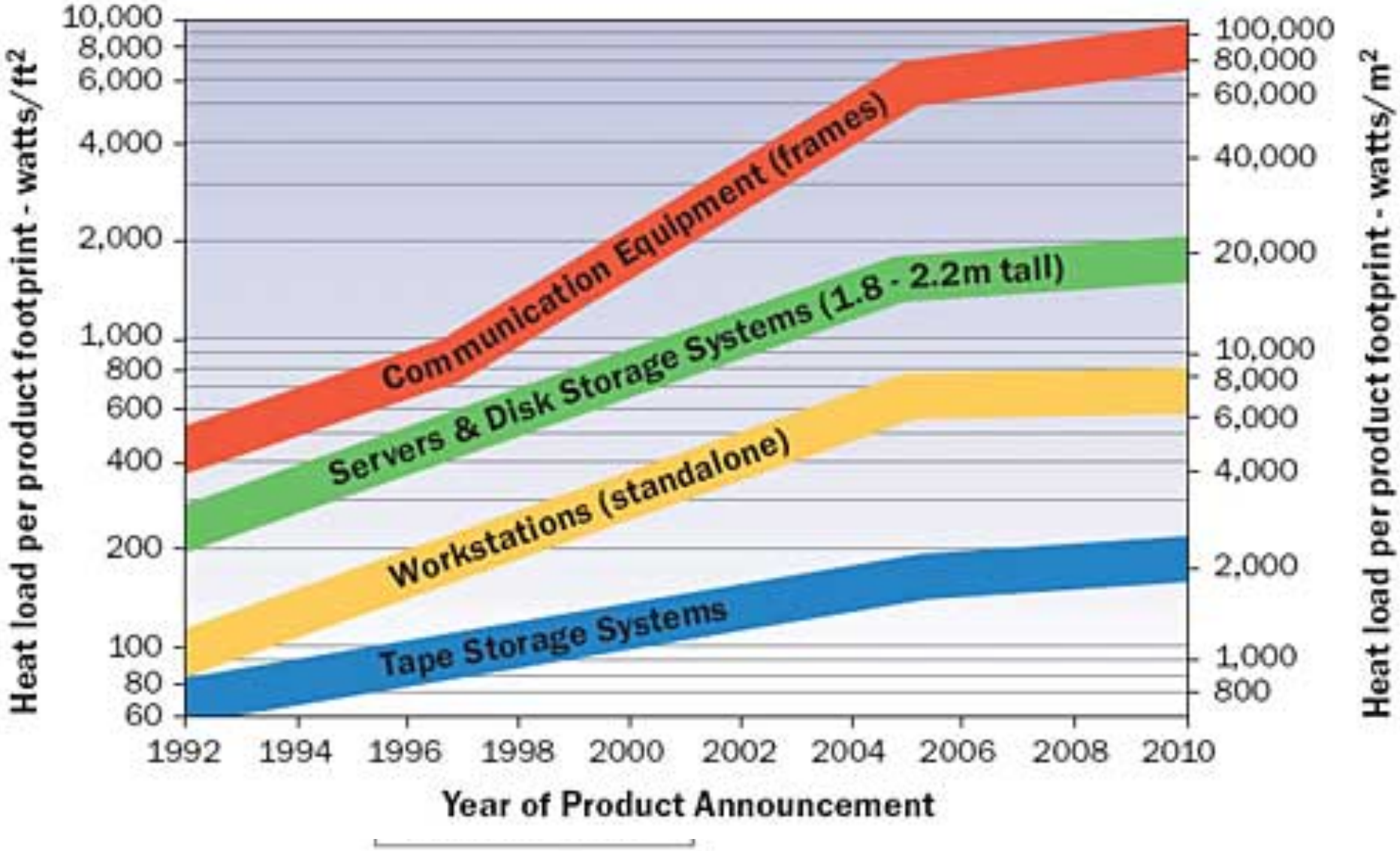
Data Link Layer / MAC: Reconsidering the MAC for new Physical layers

Thermal Image of Typical Data Centre Rack



M. K. Patterson, A. Pratt, P. Kumar,
"From UPS to Silicon: an end-to-end evaluation of datacenter efficiency", Intel Corporation

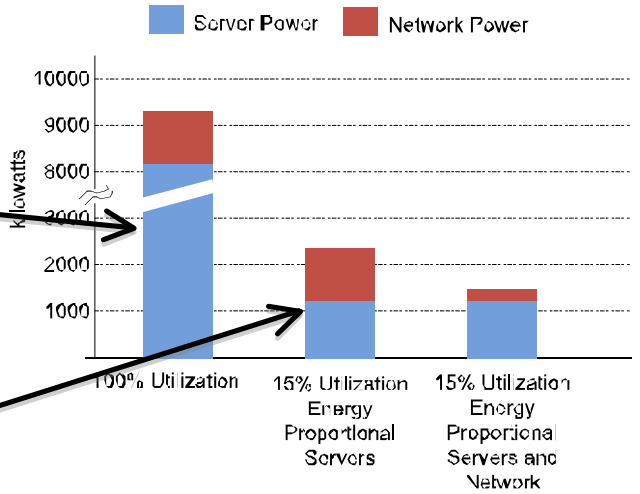
Power/Cooling Issues



Network Efficiency?

Servers at full utilization are 90% of total – particularly with improvements (h/w design and system utilization)

But an efficient server highlights the remaining inefficiencies: **the network**



Server vs Network for a Google cluster

(Energy proportional datacenter networks, Abts *et al*. 2010)

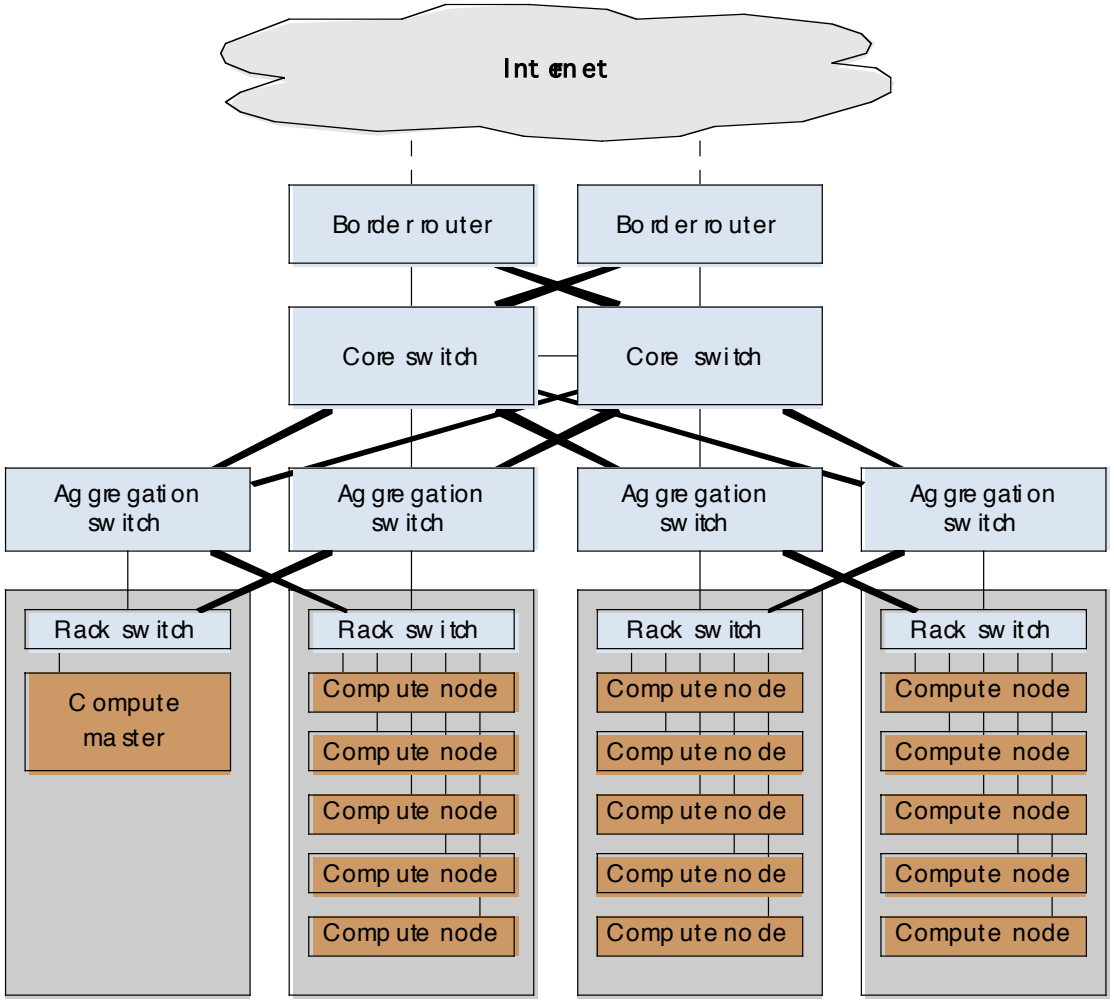
For the 15% utilization, an unimproved network may consume 50% of the total power

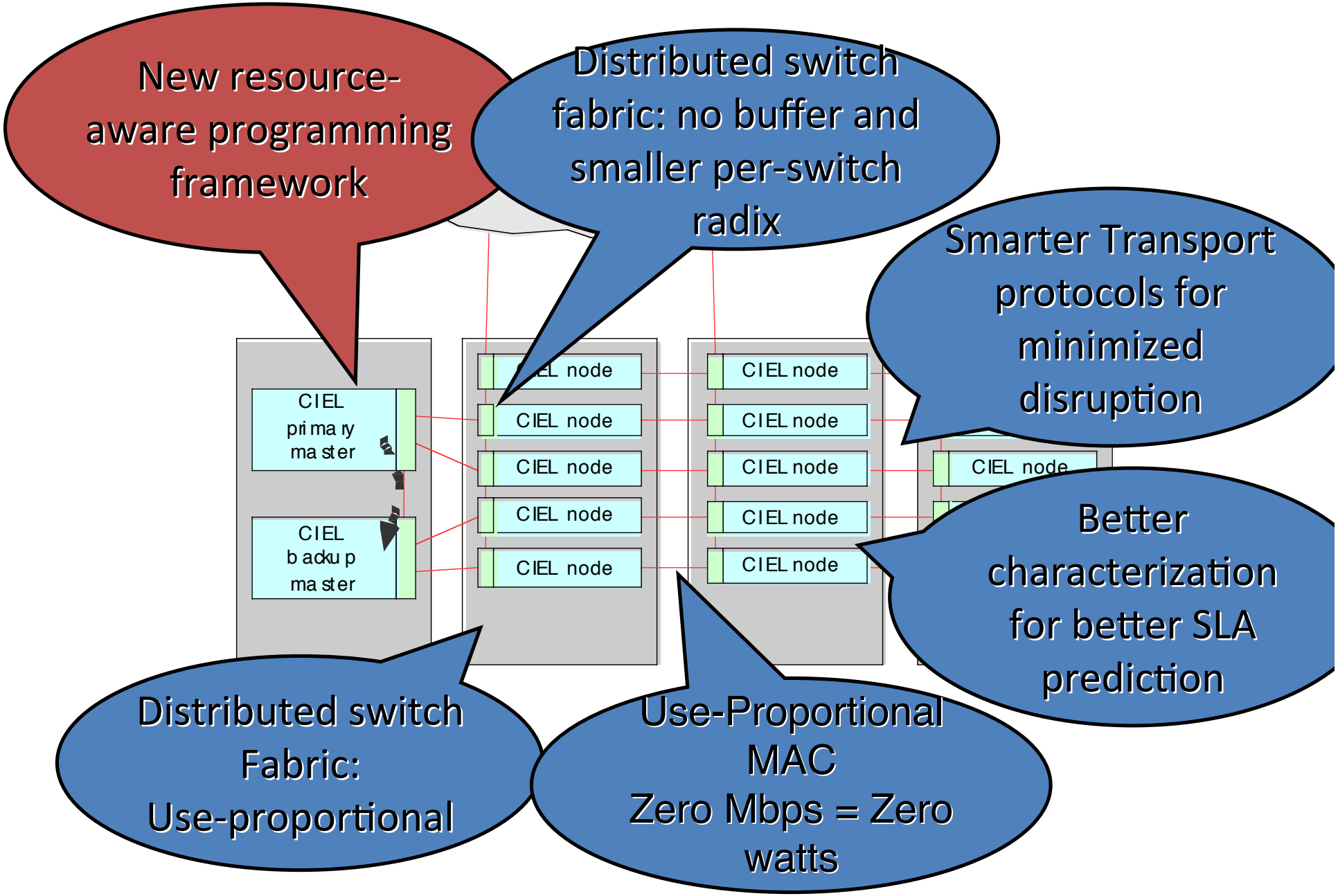
Conclusion:
We must improve the network efficiency too….

Problems

- Today's data center communications premised on multi-layer, high-performance switches
  - Inefficient/disproportionate energy use
  - Centralized points of failure
- Internet architectures are not optimal for data centers, but we use them anyway
  - Different resilience, price, performance, and security tradeoffs

A ***new*** data center approach




New resource-aware programming framework
Distributed switch fabric: no buffer and smaller per-switch radix
Smarter Transport protocols for minimized disruption
Better characterization for better SLA prediction
Use-Proportional MAC
Zero Mbps = Zero watts
Distributed switch Fabric:
Use-proportional
CIEL primary master
CIEL backup master
CIEL node

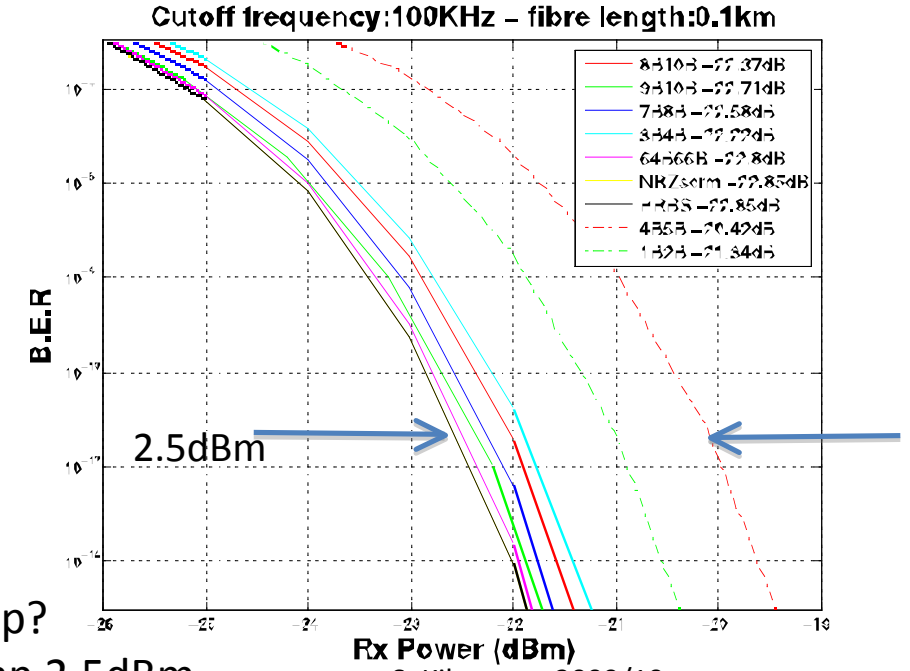
Cross-cutting themes

- Reconsidering data center switching
- Distributed resilience throughout
- Efficiency by aligning algorithm and network topology
- Energy-efficiency/security/resilience/scalability tradeoffs
- Multi-scale computing techniques

The Optical advantage

- Using optics can offer
  - Small physical dimensions
    - multiple colours can share the same fibre-path
  - Significant parallelism
    - transmitting parallel data means no delays due to marshalling (the conversion of parallel data to/from) serial data
  - Higher speeds for the same power
    - higher speeds in the electrical domain require more power, while higher speeds has no effect on power needs of optical switches
  - Distance *independence*
    - Photonics has a huge operating range (compared with copper)

It's NOT those darn lasers



S. Kilmurray 2009/10

Relatively simple modeling of the link…

Perhaps different ways of using the lasers will help?

•a range of physical coding schemes gives less than 2.5dBm

But the coding scheme consumes 4 times what the lasers consumes

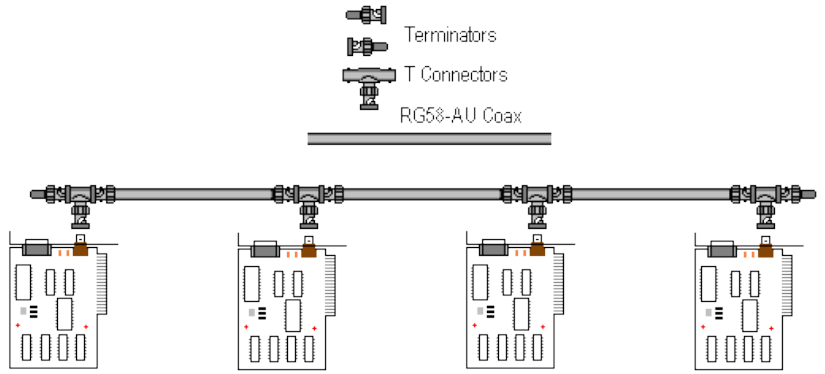
Perhaps our energy is better spent doing communications differently?

Photonic systems are better for *on demand*, that is:

better at being turned on and off as required

We need a network that works well carrying low loads, has good energy-proportionality and can quickly restart
Sounds like we need a new MAC

Remember this one? Ethernet – CSMA/CD



Many features/ideas we don't want, but one we do:

•Preambles give clocks valuable resync time
and allow photonic systems to be turned off

Preamble								Destination MAC						Source MAC						EtherType/ Size		PayLoad					CRC			
1	2	3	4	5	6	7	8	1	2	3	4	5	6	1	2	3	4	5	6	1	2						1	2	3	4

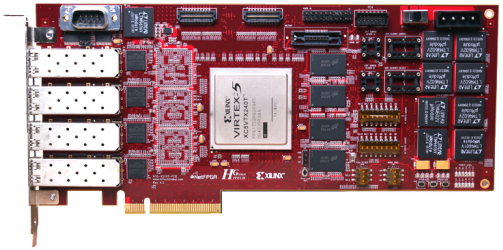
Data Link Layer / MAC

How do we test, build, trial?

- regular NICs don't help

  Need something programmable but FAST…

NetFPGA



4x10Gbps ports
PCIe x8

- Demonstrations that work at 10 – 40Gbps gets peoples attention ☺

Low Price Routers (aren't)

- Also, with Masters students looked at
  - GPUShader (KAIST)
  - RouteBricks (EPFL)
- Both use commodity hardware (GPUs or OTS PCs) as building blocks to get
  - High performance
  - Low capital price
- Both increase power usage
  - GPU a few 10s%,
  - RouteBrickes" multiples

Cross-datacentre live migration of VMs

Option 1: very large Ethernet network

- **Results in very heavy broadcast traffic**
  - 1 million hosts: >200Mb/s broadcast traffic *(Myers et al)*
- ARP, DHCP: switches can use a directory service (ELK)
- General solution: multicast
  - Automatically distinguish different uses of broadcast
  - Infer multicast groups

Option 2: migrate across IP routers

- Not currently possible: IP address changes, TCP sessions break

➢ **Need to let VMs keep IP address after migration**

- Use IPv6 auto-configuration and multi-homing
  - Small extension to hypervisor
  - No VM changes required

MADCAP

Migration-aware Data Centre Access Protocol

Toby Moncaster

(working with Jon Crowcroft, Andrew Moore)

Background and Problem

Need ability to **migrate** data centres on the fly:

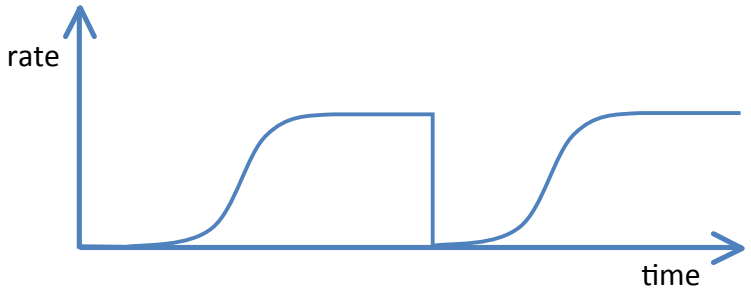
- *Preserve connections*
- *Maintain state*
- *Minimise impact on end-users*

Data centres support **multiple applications**

- *Media streaming and file serving*
- *Search*
- *Database and mail/messaging*

Most connections use **TCP** (or TCP-like variants):

- *3-way handshake*
- *Slow start - probing*
- *Significant response to time-outs*

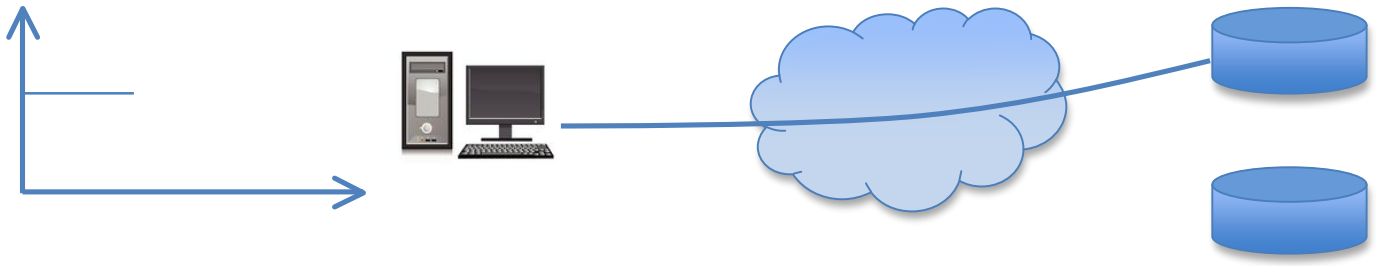


Issues to solve:

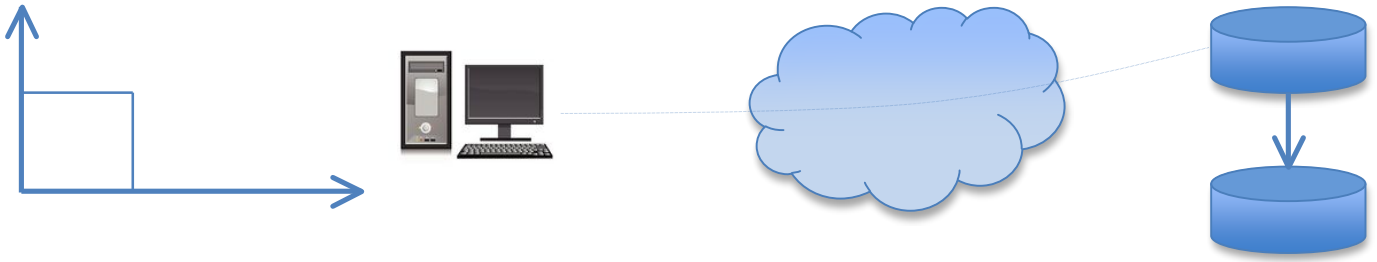
- *How to prevent TCP restarting?*
- *How to minimise delay?*
- *How to conceal process from application?*

Data Centre Migration (current)

1. Streaming data
- *Established flow*
- *Steady state*

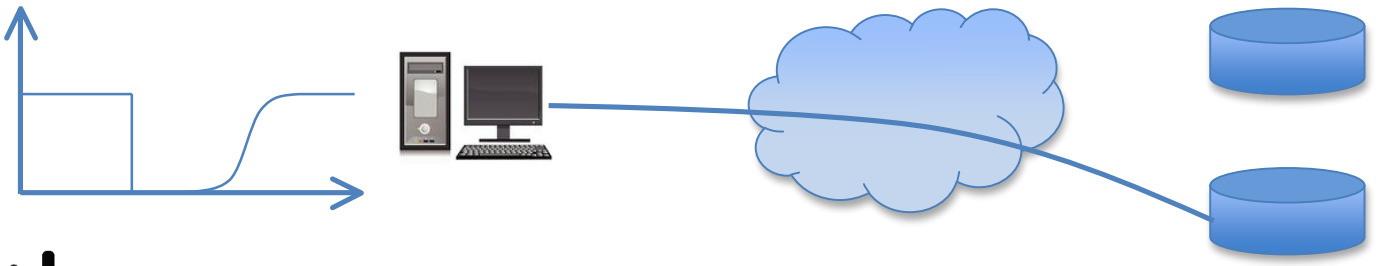


2. Migrate data centre
- *Stop flow*
- *Transfer state*



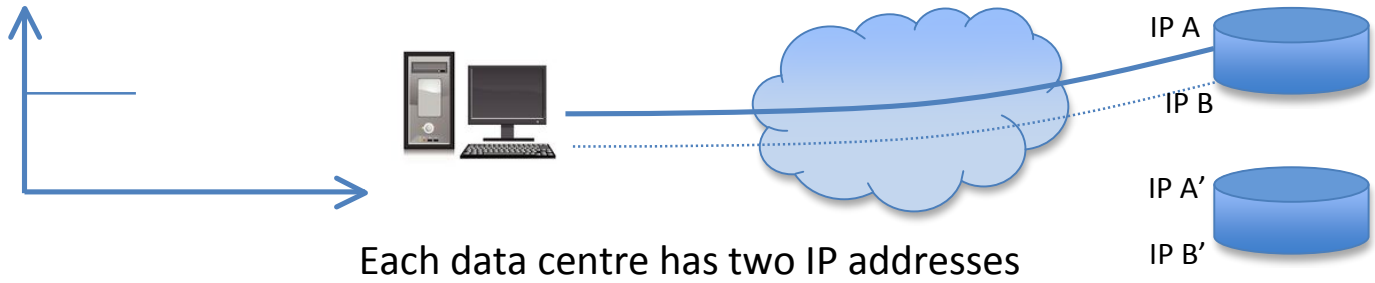
3. Restart (Best case)
- *TCP restarts*
- *Data rate increases*
- *Reaches steady state*

3. Restart (Worst case)
- *TCP never restarts*
- *Application stalls*



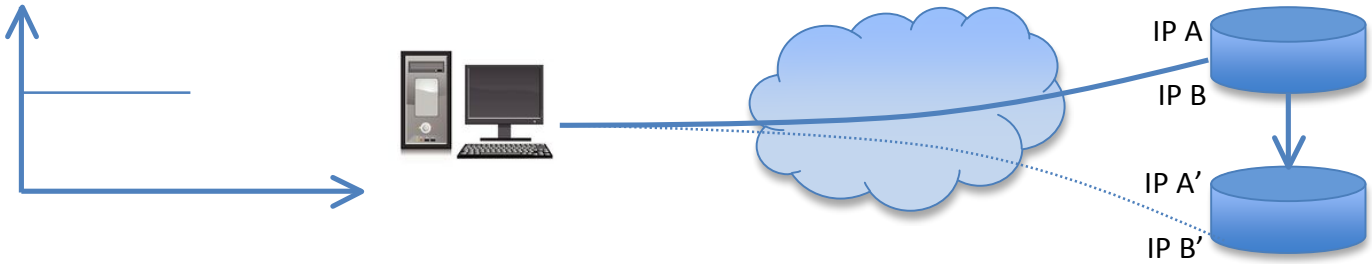
Data Centre Migration (with MPTCP)

1. Streaming data
- *Established flow*
- *Steady state*
- *Real path and shadow path*

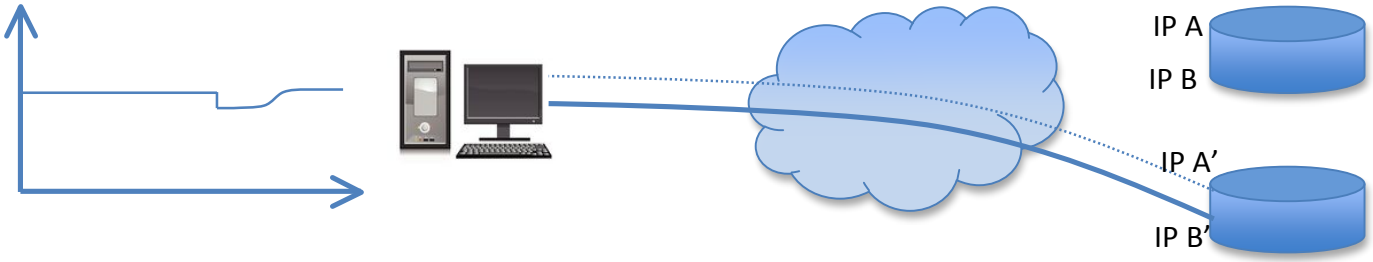


Each data centre has two IP addresses

2. Migrate data centre
- *Transfer shadow path*
- *Transfer state*
- *Transfer connection (window size, etc)*



3. Switch Flows
- *MPTCP handles transfer*
- *Application sees RTT change*
- *Application carries on as before*



Pathways to Impact

Open Standards

- Zero cost – no membership fees, no charge to use, freely available
- Everyone is equal – contributions assessed on quality
- Anyone can contribute – not a closed shop

Computer Lab has relevant experience:

- **IETF** – multiple RFCs, experience of new work groups, IAB

Open-source Software

- Zero cost – no license fees, no up-front costs
- Open community – anyone can contribute, code maintained by all
- Flexibility – open source allows you to tailor software to your needs

Computer Lab has relevant experience:

- **Xen** – significant industry buy-in, de-facto standard

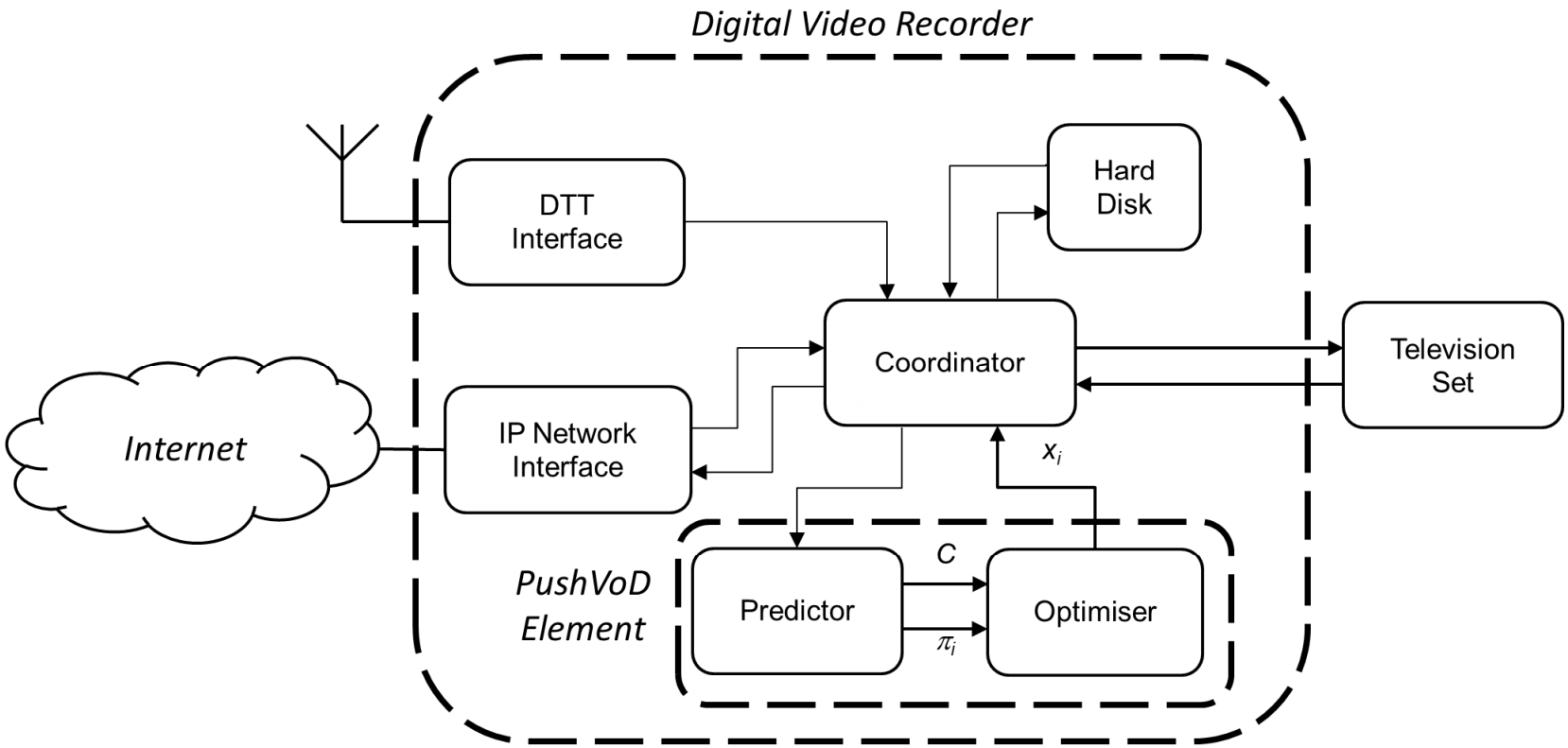
Low Energy Mobile Devices

- Talked before about ErdOS -
  - Social Operating System for smart phones
  - Shares nearby device capabilities
  - Currently working on sharing A-GPS (+map) data
  - Shows only small energy saving
  - But big speedup in TTFF (Time To First Fix)
- Also done lots on WiFi and FlashLINQ tethering

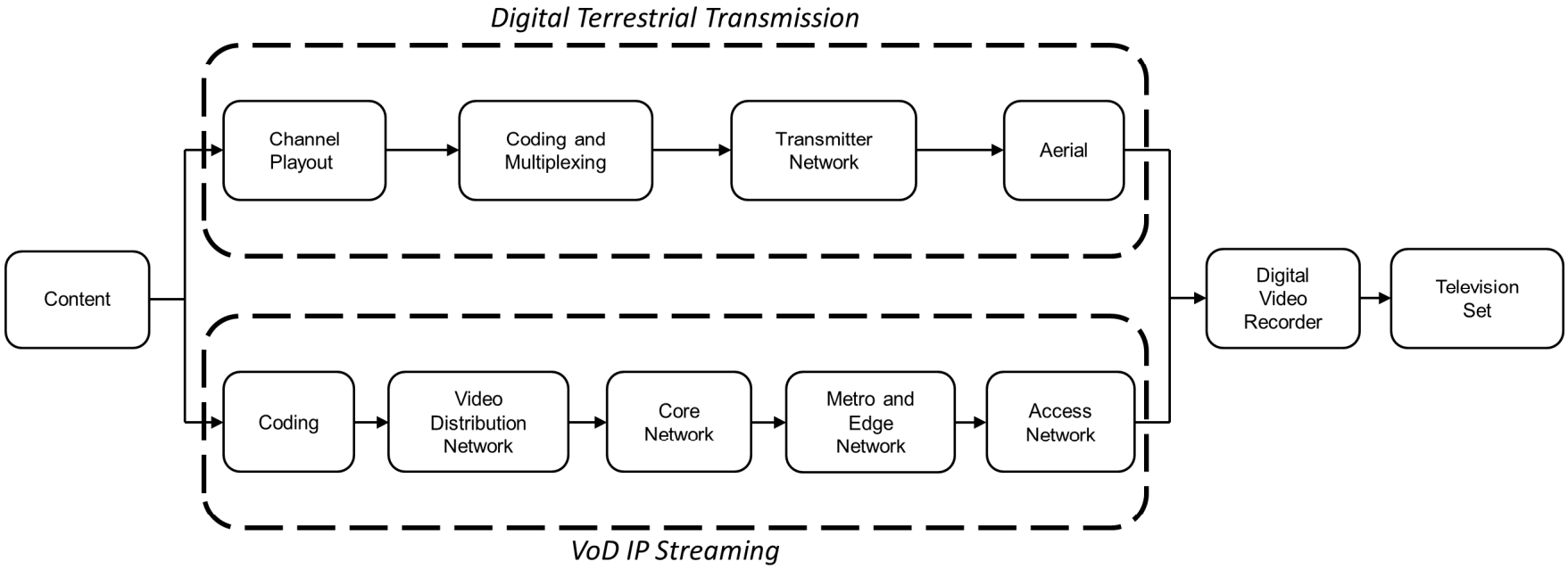
BBC Nets

- Digital Broadcast (Analog nearly all off now)
- Also carried on lots of Cable bundles (digital)
- iPlayer/Youview….

Integrated DVRs/iPlayer



Hybrid Network Delivery



BBC Stats

- Nielson-type *samples* of broadcast popularity
- **100%** detail of iPlayer statistics
  - Live Streamed or time shifted
  - Who watches what, when and where,
  - Down to house level of detail

Optimization on pushVOD

Gianfranco Nencioni

Telecommunication Networks Research Group
Dept. of Information Engineering
University of Pisa

UNIVERSITY OF CAMBRIDGE

Cambridge (UK), October 2011

Scenario

pushVOD

Hybrid set-top box (STB) automatically record content that is chosen by the content provider. When the viewer requests such content on demand, it is already available locally on the STB rather than having to be delivered via the IP network

Prediction

Determine the probability that the viewer will watch a content by basing on previous watched contents

Optimization

Choose which contents record by basing on prediction to minimize the overall energy consumption

Note: ***weekly*** *time scale of prediction and optimization*

Problem Statement - 1

Given:

- Set of possible contents: C
  - Probability of watching a content: $\pi_i \; \forall i \in C$
  - Duration of content: $\tau_i \;\; \forall i \in C$
- Power consumption of IP streaming: P^{IP}
- Power consumption of recording content on STB: P^{STB}
- Content bit rate: r
- Size of STB memory unit: S

Variables:

- $x_i = \begin{cases} 0 & \text{if content } i \text{ is not recorded} \\ 1 & \text{if content } i \text{ is recorded} \end{cases} \quad \forall i \in C$

Problem Statement - 2

Problem Formulation:

Penalty factors

minimize

$$\sum_{i \in C} \pi_i \cdot P^{IP} \cdot \tau_i \cdot (1 - x_i) + \sum_{i \in C} \alpha \cdot (2 - \pi_i) \cdot P^{STB} \cdot \tau_i \cdot x_i$$

Energy consumption of IP streaming

Energy consumption of STB recoding and watching

subject to

$$\sum_{i \in C} r \cdot \tau_i \cdot x_i \leq S$$

Memory Constraint

Problem Statement - 3

Notes:

- Penalty factors:
  - Due to the event that the user does not watch the recorded content
  - Multiplicative factor (arbitrally chosen): α
    - Maybe it can depend on the prediction accuracy
  - If no watched:
    - No energy cons. of IP streaming and watching on STB: $\pi_i \in (0,1)$
    - However, energy cons. recording on STB: $(2 - \pi_i) \in (1,2)$
- Neglecting of recording two or more contents in the same time
  - Is it a rare event?

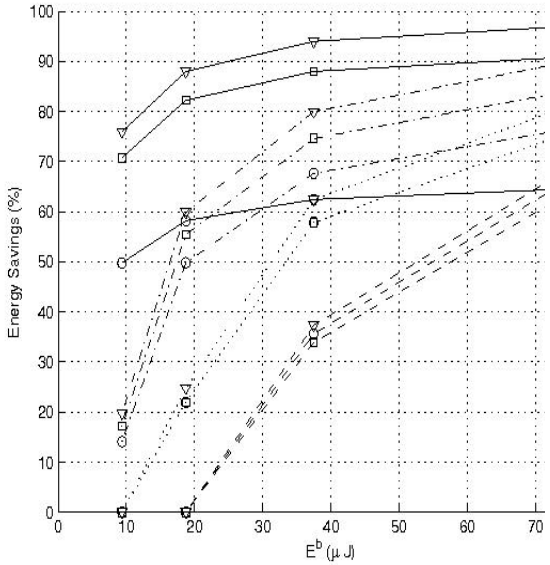
Comments

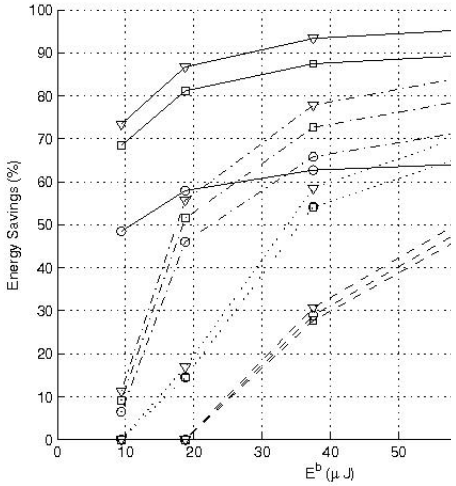
- The optimization is profitable in particular for “hungry” viewer and small memory size
- Solution of problem by means of:
  - ✓ *AMPL: algebraic modelling language for linear and nonlinear optimization problem*
  - ✓ *CPLEX: mixed-integer linear programming solver*
- Input:
  - ✓ *Predictor*
  - ✓ *Synthetic: random based on BBC traces*
- Compere optimization with choosing of contents to record:
  - ✓ *randomly*
  - ✓ *by sorting based on watching probability*

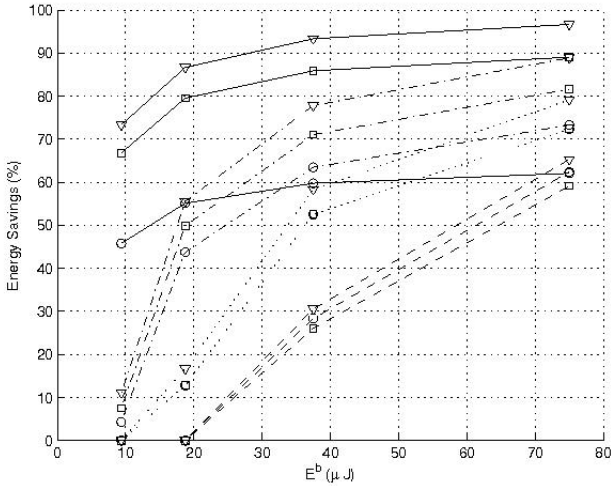
What we can optimise

- First off, predict what someone will want
  - If you watched 2 out of 3 episodes of Dr Who
  - Then you have a .66% probaility of watching next episode
- Record it when *broadcast – a.k.a PushVOD*
  - Set-top-boxes *(STB)* already measure popularity
  - Just need to integrate with iPlayer
- Model says we can get about 89% energy saving this way in theory
- Current Algorithm *only* gets 30%

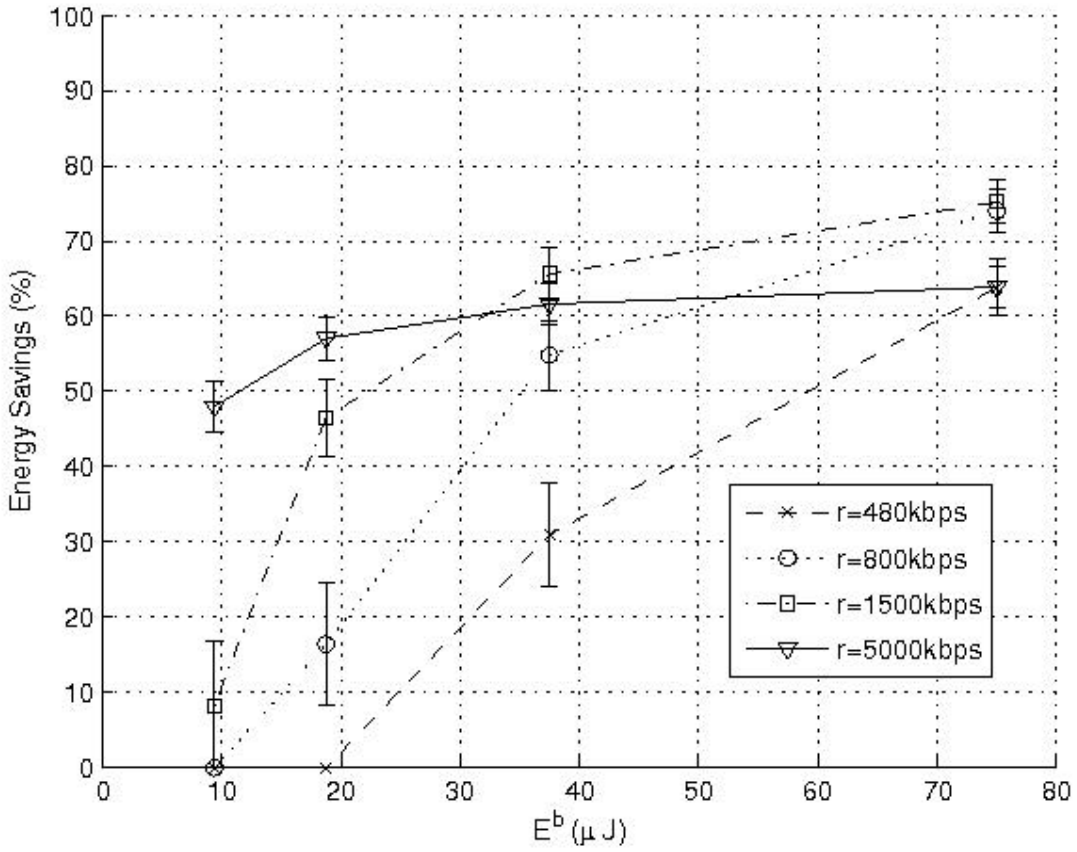
Characterizing Data over 3 different weeks



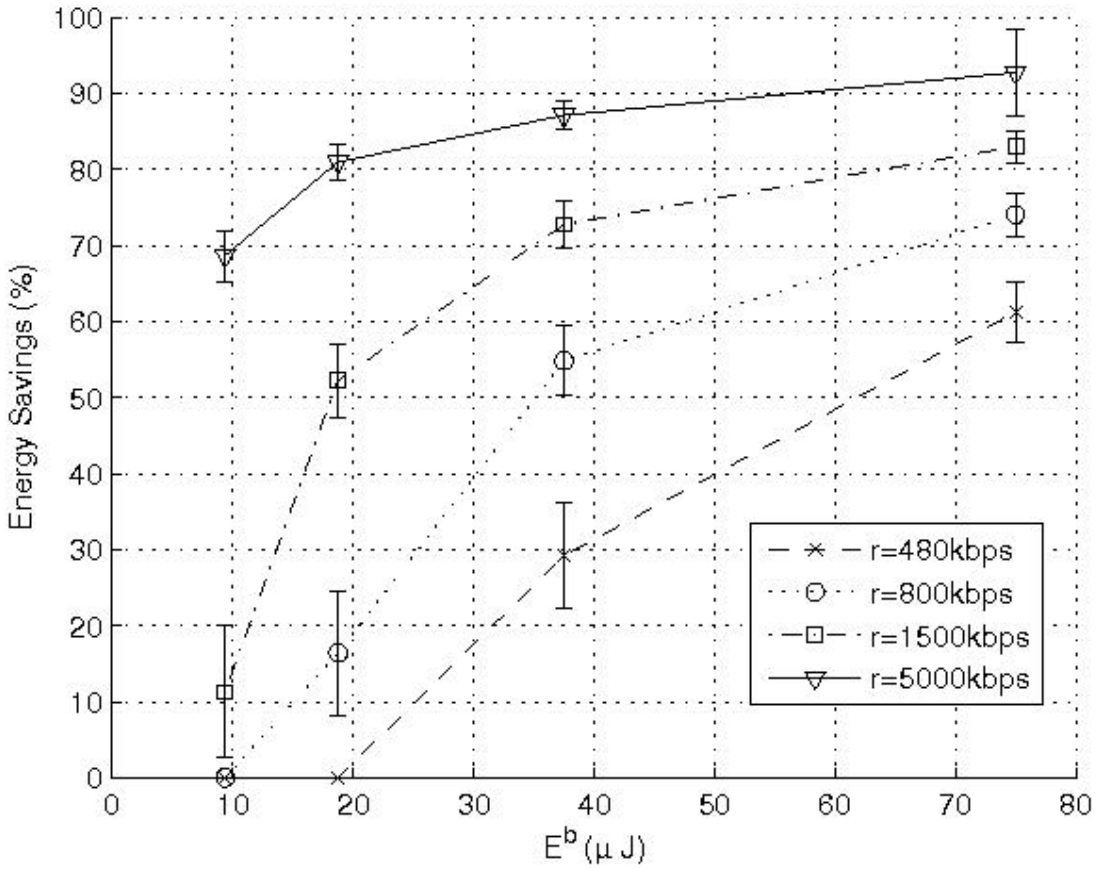




Average Energy Savings and 90\% confidence intervals - Oracle

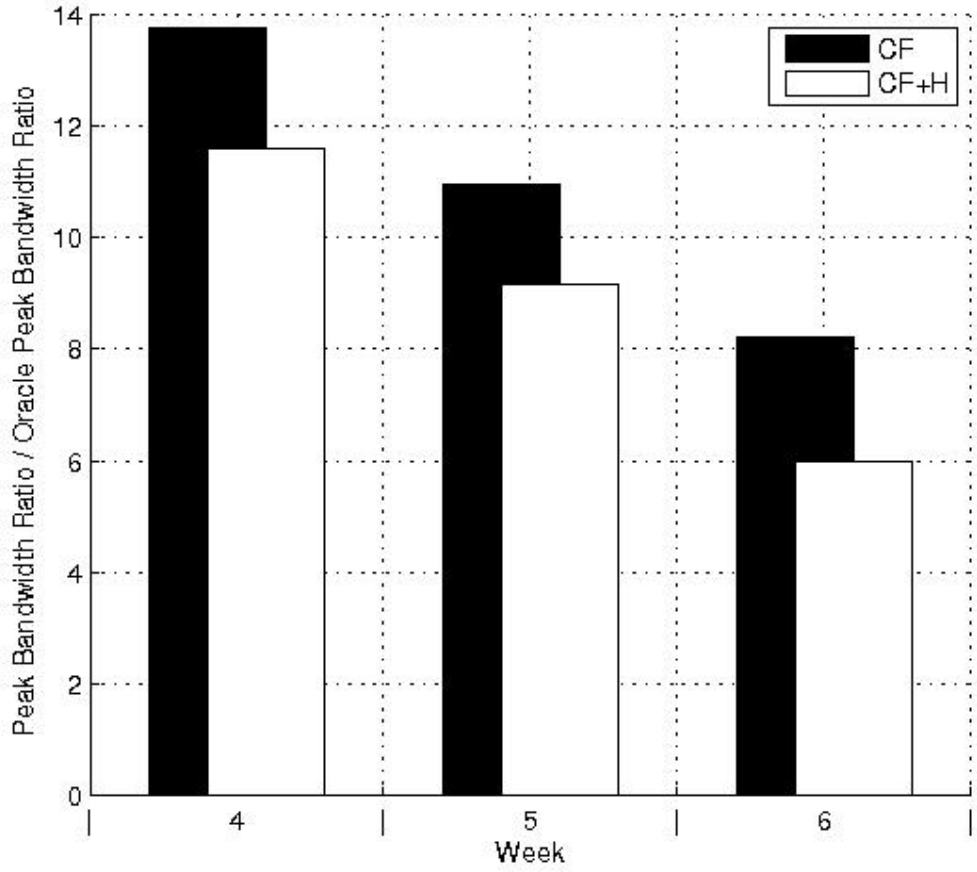


Rate Proportional S

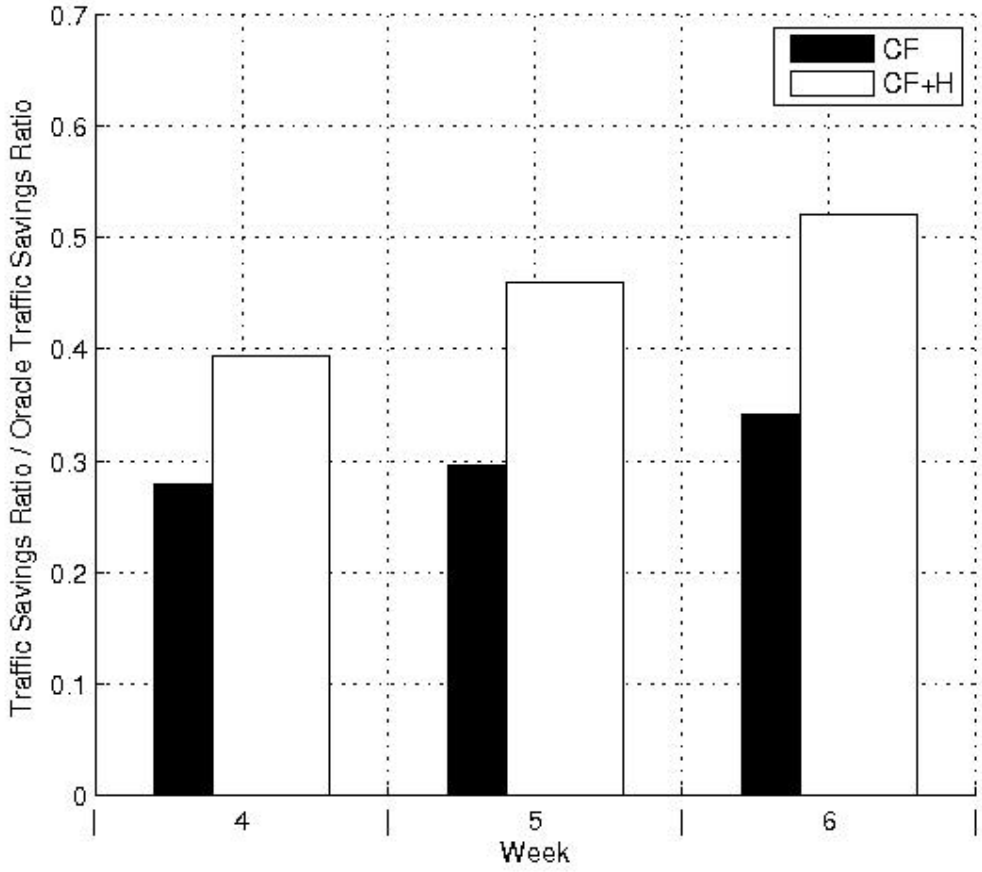


+Collaborative Filter+Heuristic

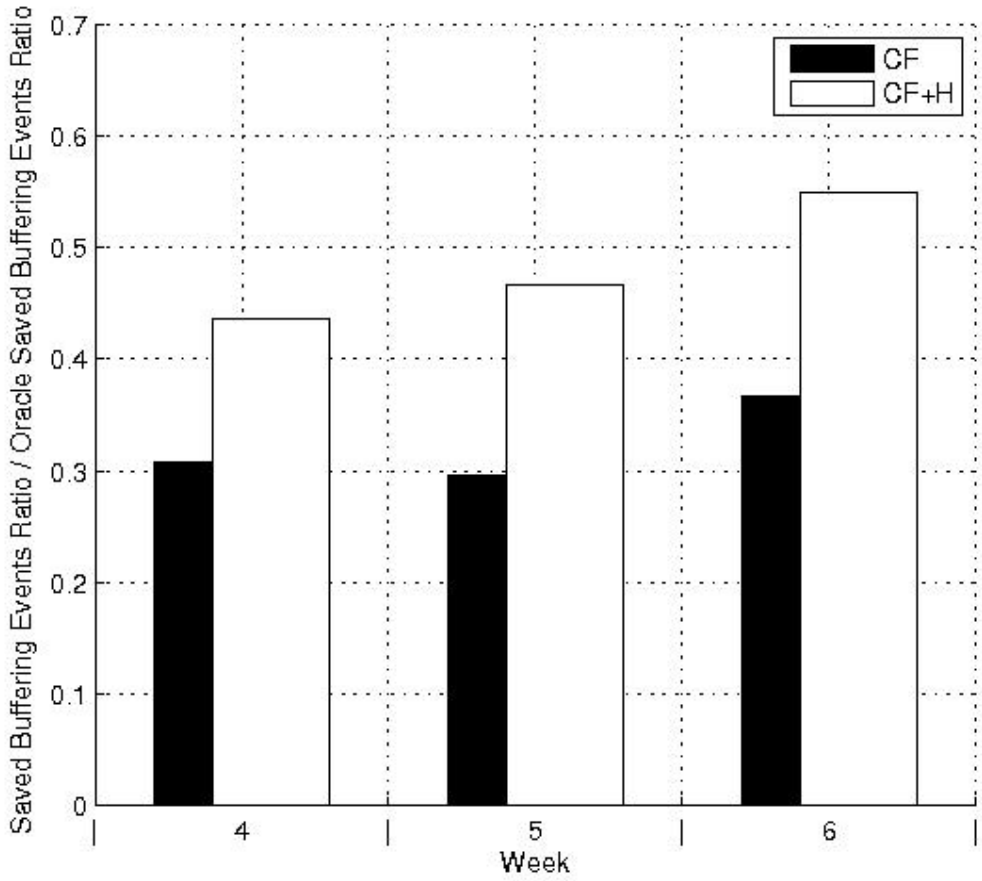
Peak Bandwidth Ratio



Traffic Savings Ratio



Saved Buffering Events Ratio



Prediction Problems

- False positives
  - Record programes that aren't watched
  - Wastes space (and power in turning on STB)
- False negatives
  - Miss programme on broadcast that is later watched
  - Wastes energy in iPlayer internet download…
- Cause is burstiness of users
  - Need longer estimation window to refine prediction

Ironically…

- BBC want to turn off digital broadcast
  - Having just turned off analog broadcast
- Need replacement
  - Could do IPTV multicast
  - Like AT&T and Telefonica.
- Could also look at swarms
  - Bittorrent (resource pooling) known to be optimal
  - Cf. Akamai doing now (mix of CDN&P2P) for IPTV
  - Need to re-run Energy Analysis/Models

Conclusions

- Optics
  - Biggest opportunity, longest timeframe
- Migration
  - Useful future - unevenly distributed
- Network Optimisation
  - Can do now - many ways
  - BBC specific example

Q&A