Energy Efficient Ethernet Beyond the PHY

Power savings in networked systems

Rudy Klecka, Hugh Barrass (Cisco)

A wider view

EEE study group has discussed saving power in the PHY

- RPS – or similar mechanisms

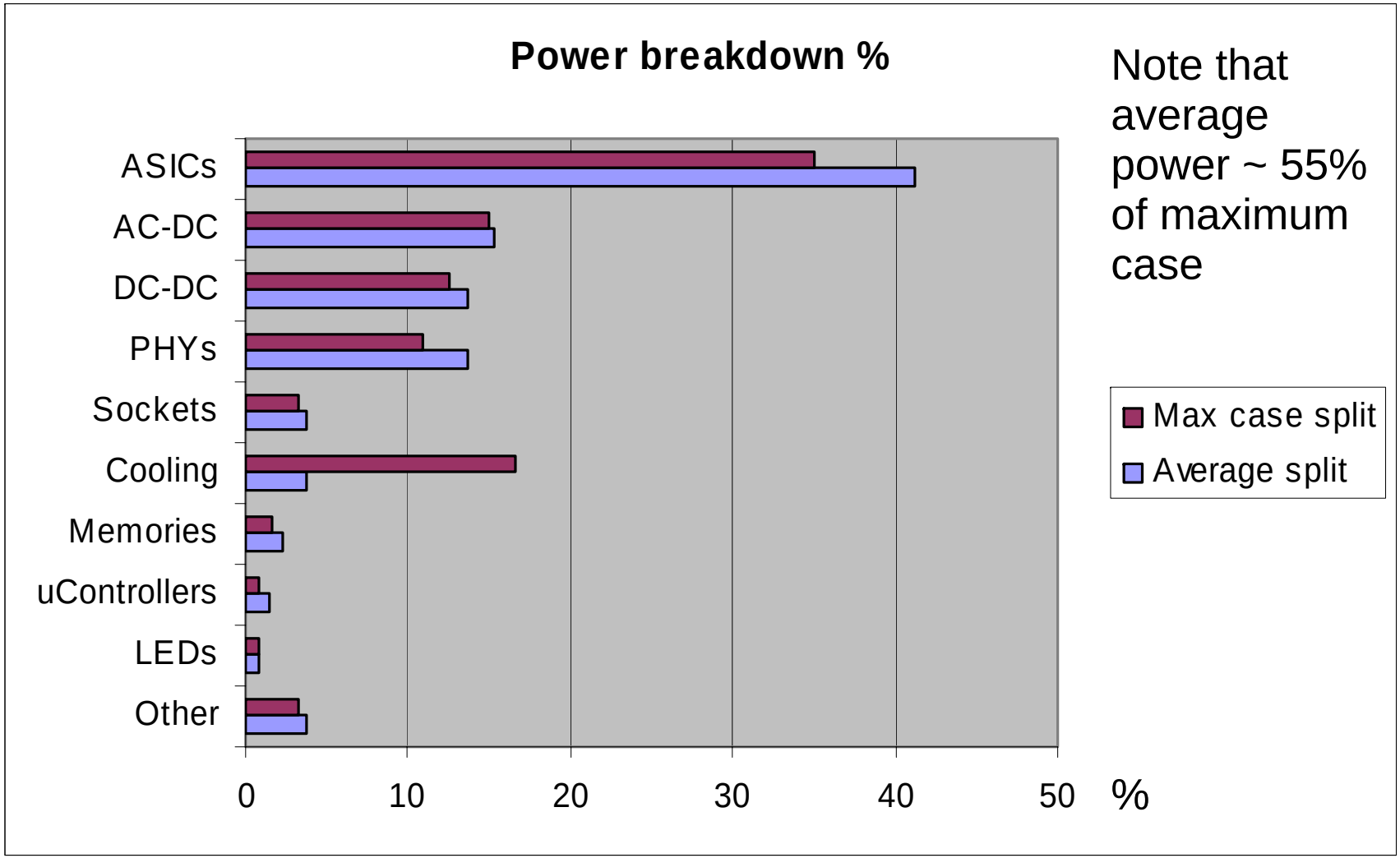
But whole system power measurements shown in CFI

- Power savings vs PHY speed > expected PHY power
- Even existing systems are saving more than PHY power

Examine current and potential system power savings

- “Reduction of power during low link-utilization”
- Where will this benefit from standards-based control?

Where does the system power go?



Note that average power ~ 55% of maximum case

General design issues

- AC-DC & DC-DC conversion efficiency
  - Many existing supplies lower efficiency at partial load
  - Next generation supplies improve efficiency curves
- Variable cooling with load is becoming common
  - Reduces 24/7 energy use considerably
- Some systems are lowering power of “sundries” with usage
  - e.g. system uController, management etc.
- Designers are choosing static vs dynamic wherever possible
  - Lowers typical power consumption
  - May include static data for idle states
- None of this requires interoperability standards

PHY power savings

Assuming a mechanism such as RPS…

… and assuming it works as required ☺

- PHY power saving is straightforward and easily understood
- For Cu PHYs: 10G -> 1G ~ 4W (in future); 1G -> 100M ~ 0.5W
- Also, newer 10M PHYs could be ~0 power for 0 traffic

This would be significant for typical 1G & 10G systems

- ~12% at 1G; ~30% at 10G – compared to typical power usage

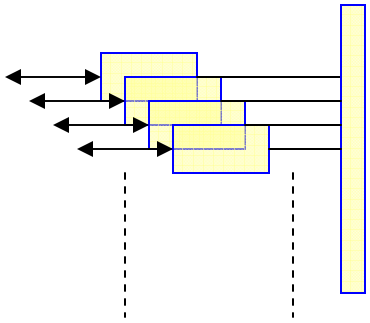
This power saving would need a new standard

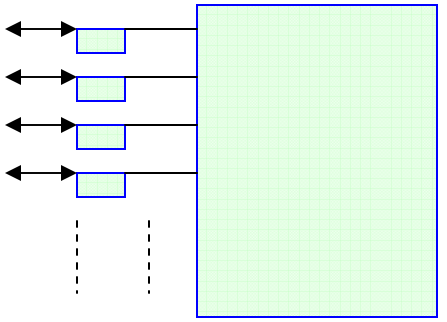
- Something like RPS – both ends of the link
- Hence the focus in the CFI…

ASIC power savings

- Power dominated by clock speed (for typical processes)
  - ~50% for current, >60% in future
- ASIC power savings highly dependent on architecture
  - Heavily port-based architectures can work with link speeds
  - Centralized fabrics more difficult to deal with
- Data activity (or lack of it) saves power automatically
- Other savings require conscious effort
  - Change clock speed or data width
  - Switch off power to “islands” within the chip / system
- Time to change, buffering is critical, + policy vs architecture

ASIC architectures





Port-based

Distributed memory structures & data paths

Easy to reduce single port, link speed or RPS

Power savings smaller – law of small numbers

If memory structures used to absorb "return to activity" burst, cannot be powered down

Centralized

Large central memories & data paths

Power saving modes depend on thresholds

Large power savings % for v. low activity

May require port memories to absorb bursts, hinders efficiency

Traffic characteristics = aggregation

NB – power savings in ASIC memory structures discussed later

ASIC power savings

Existing architectures save power with link speed

- Static functions, based around auto-neg speed
- Saves 0 – 25% of ASIC power (up to 10% of system)
- Could benefit from RPS without architecture change

Newer architectures could increase savings

- If known changeover time before high speed burst…
- … allow more widespread shutdown (without buffer wastage)

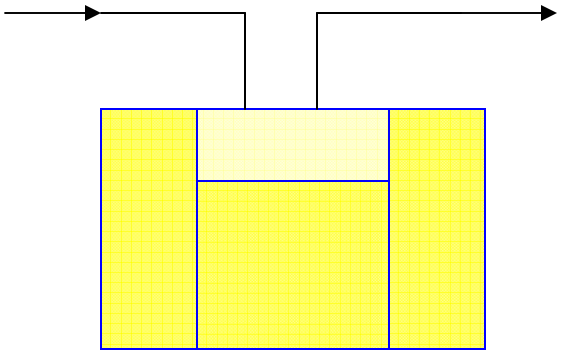
High speed, centralized designs could save >75% of ASIC

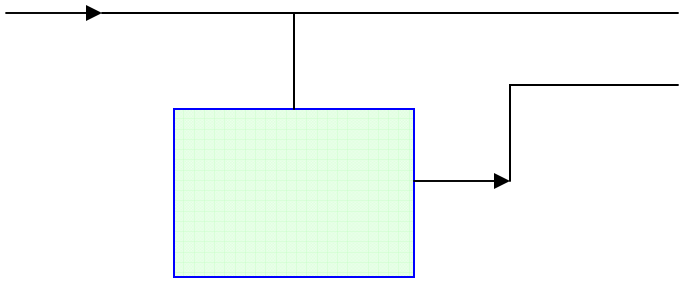
- Up to 40% of system power
- Controlled speed change allows larger savings…
- … uncontrolled requires more conservative policy

Memories and ASIC memory structures

- 2 types of memories considered: data path storage & lookup
  - Data path memories are often dynamic – v. wide, v. fast
  - Lookup memories may be CAMs, hash tables, etc.
  - Both external and embedded memories (embedded counted in ASIC)
- Static memory power savings largely automatic
- Dynamic data path memory savings
  - Reduce clock speed with link speed
  - Reduce depth and/or width for access & refresh
- Lookup memory power savings
  - Reduce clock speed & width
  - May require longer change-back time to reload (~ 100's uSec)
  - Larger potential savings in typical architecture

Memory types





Store-and-forward memory

Some savings come from reducing width

Bandwidth.delay sizing allows depth reduction, saving in exercising columns and refreshing

Lookup memory

Reduction in width for lower bandwidth

Some architectures may eliminate parallel copies for high bandwidth support

Dynamic copies may need reload for “return to activity”

Static memory power scales with activity

SSRAM benefits from reduction in clock speed or width

In conclusion

Current systems already scale energy use with activity

- 40% reduction from peak to typical (10% load)
- Much smaller reductions from typical to light load

Next generation architecture should offer more savings

- Beyond PHY power, savings >50% of typical

~ half of savings achievable through good design practices

- Remainder require open standard
- Reliable link partner behavior, known changeover time

Secondary benefits of standard, also helps energy savings

- Common behavioral expectations, clearer benchmarking

Questions…

… or comments