Energy Efficiency Challenges for Exascale Computing

Alan Gara
IBM Research

© 2008 IBM Corporation

Outline

- Power issues under analysis for BlueGene
- Where we are going long term.. (targets)
- Power scaling from 20 PF/s in 2012 to 1000 PF/s in 2018
- Key technology plays
- Key architectural directions and mindset issues

The Green500

- Improvements in system power efficiency have been marginal lately.
  - IBM Blue Gene/L in 2005 210 MF/W
  - IBM Blue Gene/P in 2007 370 MF/W
  - IBM cell based systems in 2008 480 MF/W
  - **Factor of 2.5 over 3 years. Constant power/system would require a factor of 8**

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)	TOP500 Rank*
1	488.14	IBM Germany	BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband	22.76	324
1	488.14	Fraunhofer ITWM	BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband	18.97	464
3	437.43	DOE/NNSA/LANL	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz , Voltaire Infiniband	2345.5	1
4	371.75	Argonne National Laboratory	Blue Gene/P Solution	31.5	304
4	371.75	Dublin Institute for Advanced Studies/ICHEC	Blue Gene/P Solution	31.5	305
4	371.75	Science and Technology Facilities Council - Daresbury Laboratory	Blue Gene/P Solution	31.5	306
7	371.67	RZG/Max-Planck-Gesellschaft MPI/IPP	Blue Gene/P Solution	94.5	52
7	371.67	Stony Brook/BNL, New York Center for Computational Sciences	Blue Gene/P Solution	63	75
7	371.67	ASTRON/University Groningen	Blue Gene/P Solution	94.5	51
7	371.67	IBM - Rochester	Blue Gene/P Solution	126	37
7	371.67	DOE/Oak Ridge National Laboratory	Blue Gene/P Solution	63	74

June 2008 Green 500 List:
If Worlds' most power efficient is extrapolated to a sustained Exaflop, power would be …

~ **2-3 GigaWatts**

The Big Leap from Petaflops to Exaflops

- We just hit a Petaflop…. We are now beginning to do the research to architect for the Exascale.
  - Much of this involves deep dives into emerging technologies and anticipating the scalability of current technologies.
- Technology disruptions in many areas.
  - Silicon power scaling, Interconnect bandwidth, Memory technology, Packaging technology.
- CMOS generations are trying to hit every 2 years. We are trying to double performance every year.
- Need to be able to exploit machines. Not just about flops. Flop metric promises to be an even poorer predictor of sustained performance.
- Targeting to reach an Exaflop in 2018 at 20 MWatts

The Power Problem (Many factors contribute)

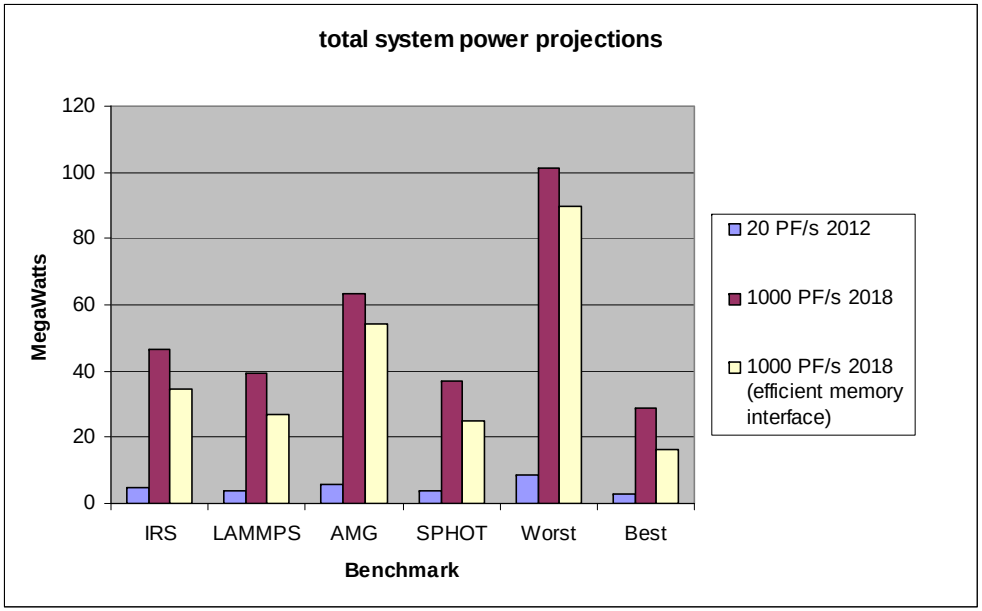
- To scale up everything needs to scale. (Looking for about 100x power efficiency improvement from current best, 10x from 20 PF/s design point)
  - Processor Core
    - Circuit technology alone will fall 4-10x short in 2018. Architecture will help with this.
  - Floating Point Unit
    - Not the biggest problem. This is probably only a `2x problem in 2018. We are already at 60GF/s/W in 45nm. This corresponds to about 250GF/s/W in 2018 for FPU power . 1EF/s of FPU alone is then 4MW
  - Cache Power (SRAM)
    - SRAM has different scaling properties to logic. Voltage and distance need to be carefully considered. If you want to utilize the larger silicon to build bigger caches, the power will only scale approximately as V^2 (V will be about 1.5-2x as compared to 20 PF/s design point)
  - Connections between different cache levels
    - Wiring will scale as V^2. Distance reduction is critical. V^2 will get us about 3-4x over 20 PF/s design point.
  - Main memory connection
    - Very difficult problem with existing packaging solutions. The limited numbers of I/Os results in using high power, high speed techniques. As bandwidth requirements grow, interfaces tend to go toward high speed data capture techniques with compensation making power usage nearly uniform. Very bad for typical applications.
  - Main memory chip power
    - Distance will be the same issue. If we expect to have many more DRAM bits because of density improvement, then we will not get the power savings from shorter path lengths and the power will only improve by V^2
  - Network power
    - Electrical still the lowest power solution. Optics has the same electrical interface so the power tends to be largely additive. High speed link power is only anticipated to improve 2-5x between now and 2018. Need to drive the voltage down to get better power efficiency but cost gets in the way of this.
  - Disk power
    - Disk power has been largely constant for each drive while spinning. Commodity driven constraint. Performance improving very slowly with continued density improvements although slowing. Should only expect ~4-10x performance improvement in spinning disk power efficiency (GB/s/W).

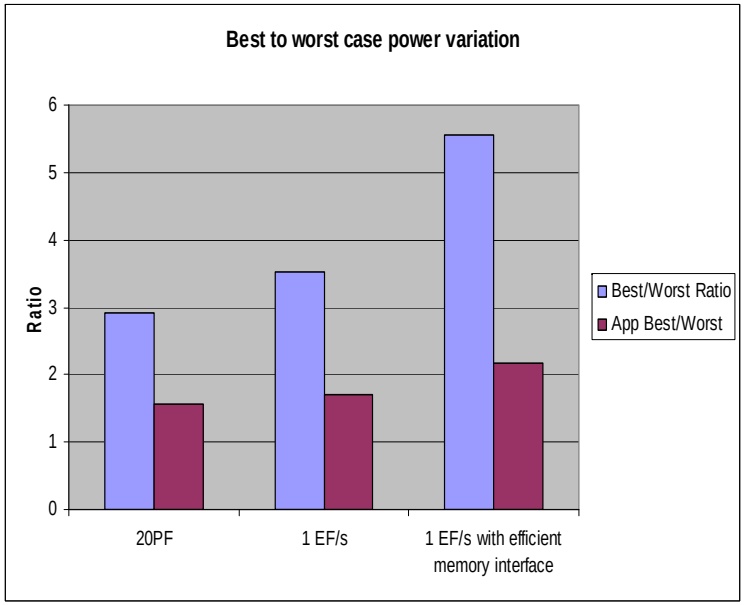
Power = c v^2 f

Power efficiency = c v^2

Scaling results to Exascale (1)

- Scaling from what we believe is possible in 2012 to an Exaflop
  - Scaling assume algorithms will have little or no improvement in terms of memory locality.



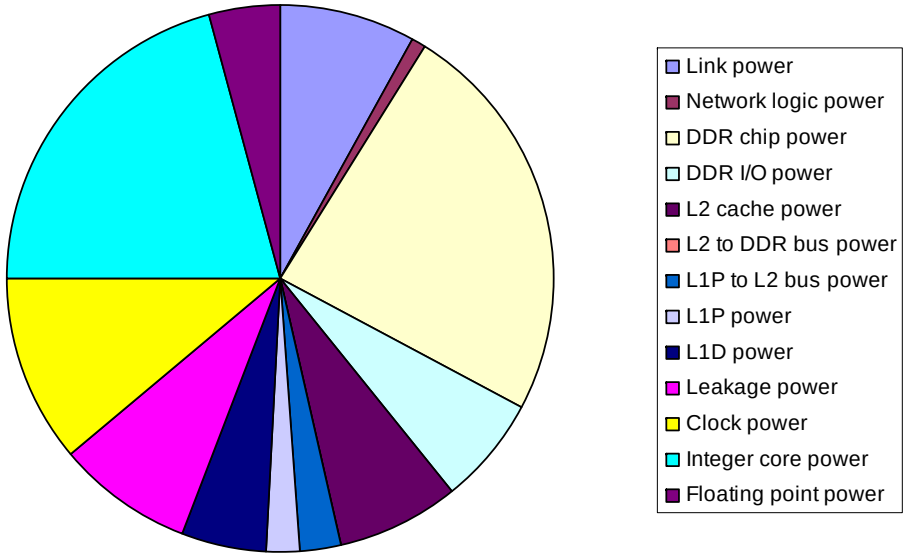


- 1 Exaflop in 2018 will consume much more power unless there are many system level technology improvements
  - Improved memory interface has a strong impact on power.
- Variation in power consumed will increase at the Exascale. (probably underestimated here)
- Worst case design point would drive almost 40MW of wasted power infrastructure

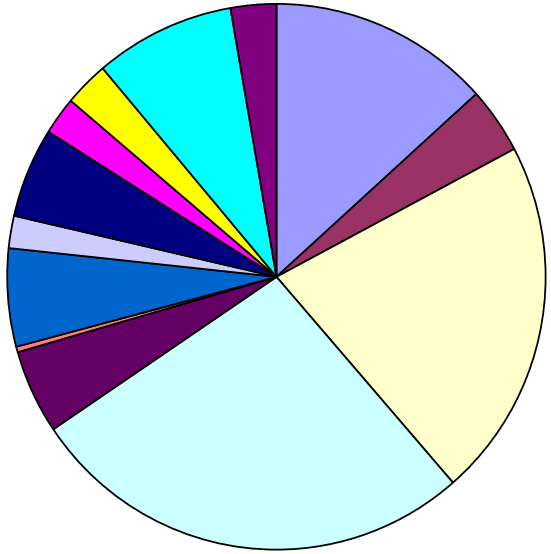
Scaling results to Exascale (2)

- Memory power and I/O to Memory are biggest power problems at the Exascale

20 PF/s 2012 typical application power distribution



1 EF/s 2018 typical application power distribution

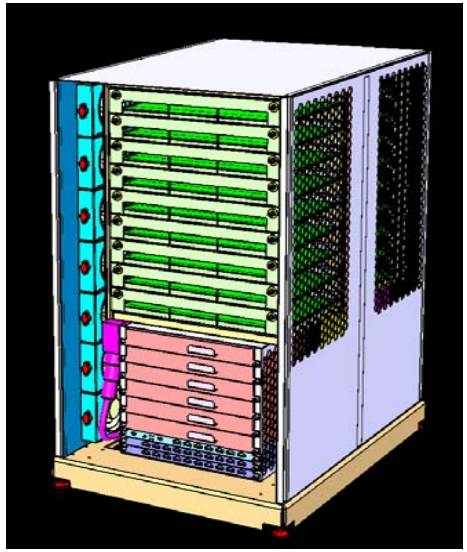


Key Technology From a power perspective : Memory

- Memory power is on track to dominate the power for Exascale systems.
- Need to aggressively interact with memory manufacturers.
  - Need more focus on the power problem
  - Memory has traditionally been focused on two main components:
    - 1) Best cost/bit (highest density beans long bit lines, higher power/bit architectures
    - 2) High speed graphics memory (poorer Perf/Watt than high density)
- Emerging memory technologies offer a chance to place perf/Watt at the right level of importance.
- Caches need to be optimized for power as a first priority rather than performance. High hit rates are very attractive but high power, low level caches are currently very power inefficient.

Key Technology From a power perspective : Packaging

- Traditional packaging solutions will not scale to Exascale. The power in the I/Os to the memory system for example would drive 80 MWatts. This is fundamental physics.
- The number of pins that are needed for node to node with naïve scaling is about 5k per node (at 20 Gbps) which is not viable.
- Need to move to a tighter integration with memory. This will require at least a 10x drop in capacitance. 3-D and silicon-on-silicon does offer such promise.
- Local density issues can be handled by simply reducing the density of processing but while this will not likely have a dramatic effect on usage modes, it will have a dramatic effect on cost as optics between racks are related to surface to volume.
- Packaging changes need to be supported by multiple vendors… Need standards.



System level throttling and wasted power

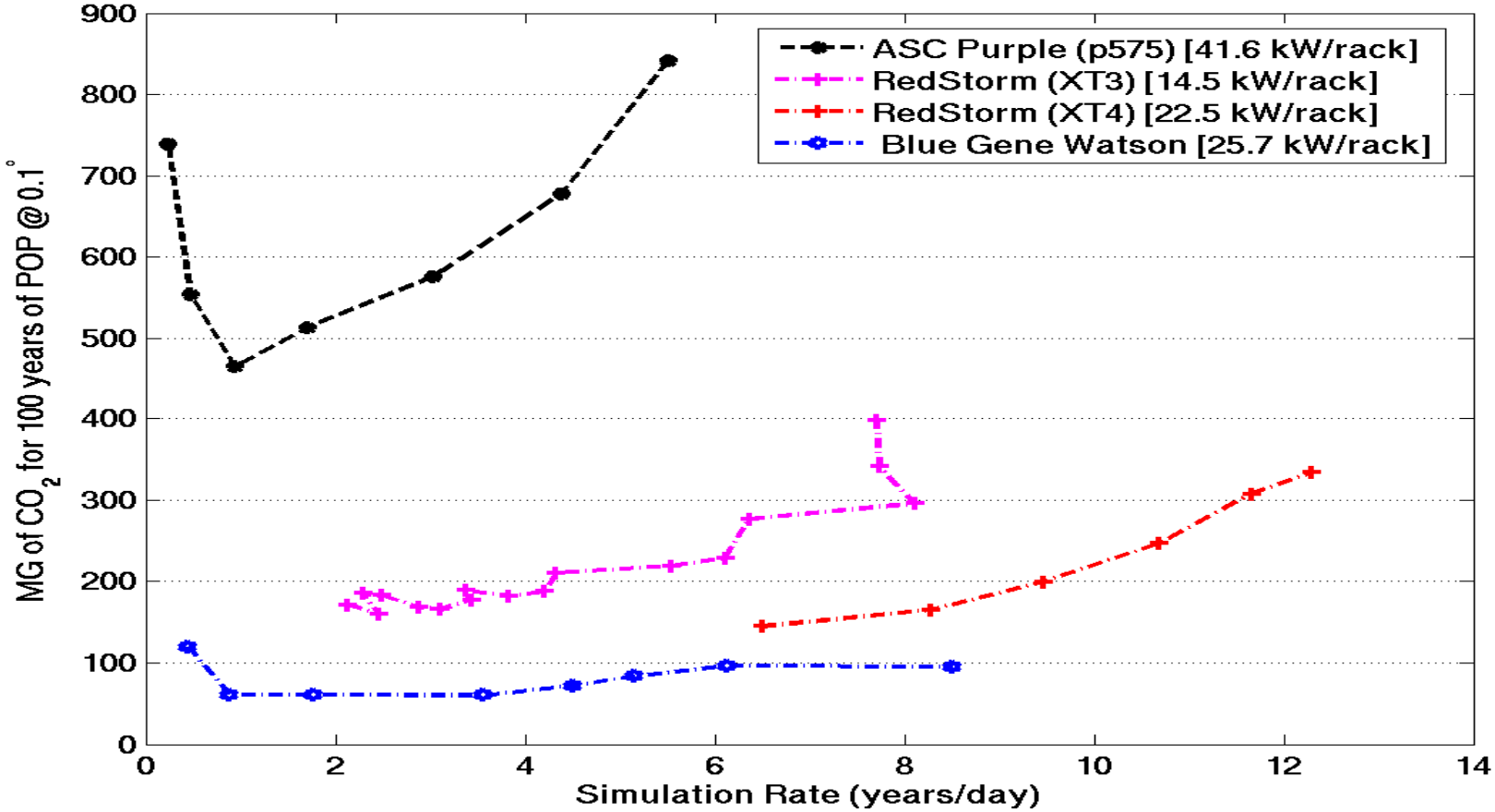
- Throttling
  - Need to accept power as a limit to performance.
    - Designing for worst case power is very cost (and power) inefficient.
    - Need to do this in a manner that is consistent with reproducible performance.
    - Power usage as the most appropriate cost metric will likely emerge
  - Tools and architecture support for power optimized system utilization are needed.

- Wasted power
  - The majority of the power in current systems is not doing useful work.
    - This is good news as there is still room to improve power efficiency through architecture.
    - Other areas where improvement can be engineered (power supplies, cooling efficiency etc)
  - Need more holistic approach toward handling power. Chip, cards, racks, facilities

Total cost , who pays. Mindset acceptance of power as the correct currency. Real green impact

- Total cost is current “best case” impact on power.
  - Cost = Machine cost + $$/MW * Power_load * time
  - It is hard to imagine this continuing as the impact to power.
  - Expect a true “green perspective” on power to emerge. (more than just the immediate $ impact)
- Power impact not always included in total cost as facility costs are longer term and handled separate from machine acquisitions.

Carbon footprint for simulation of POP@0.1°



Courtesy of John Dennis

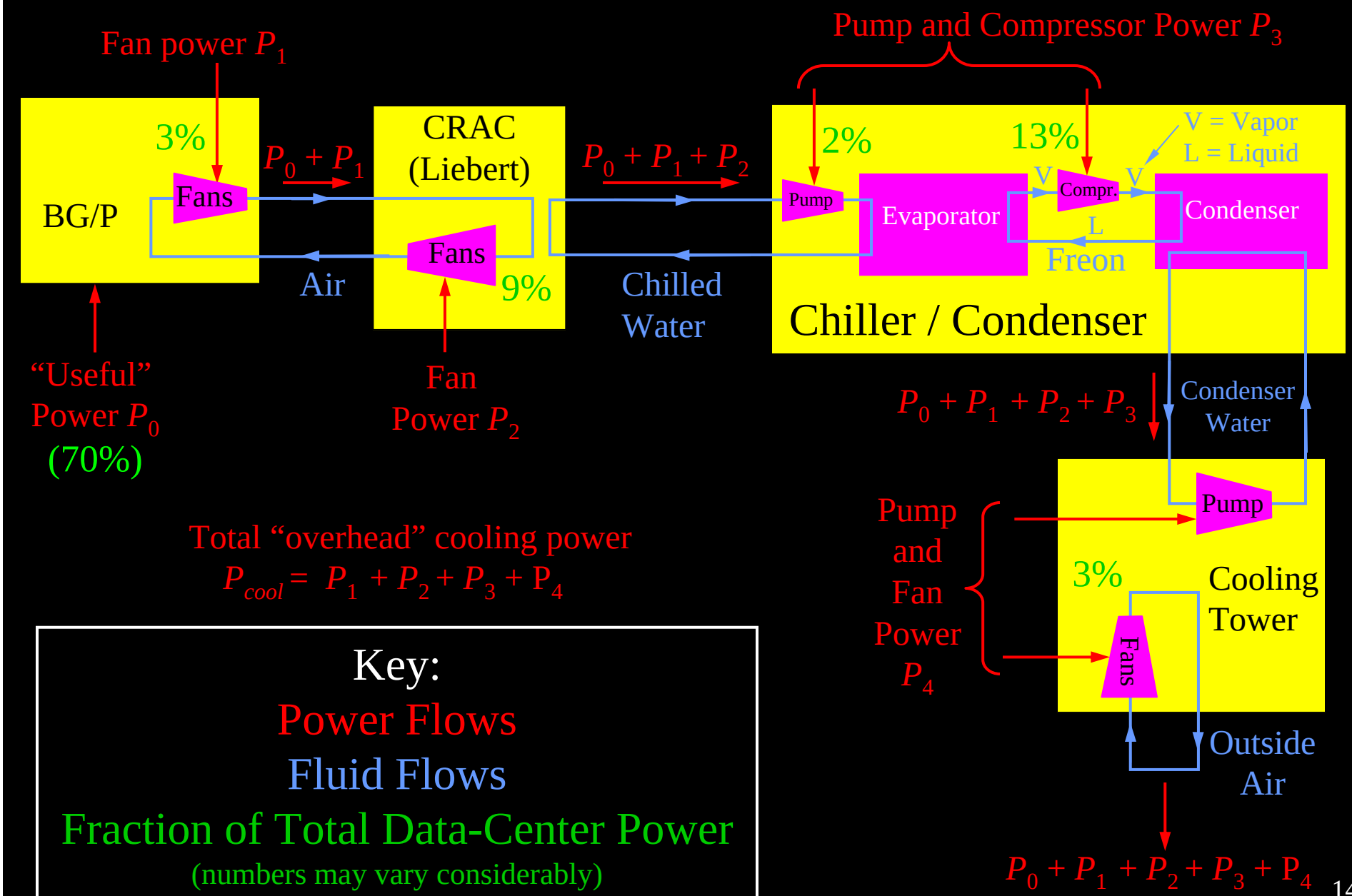
Reliability

- Reliability and Usability need to be maintained as we move toward the Exascale.
  - Unlikely that we will be productive if we leave this to users.
  - Power will drive the industry to lower voltage where problems get worse.
- New technologies can be leveraged to deliver a reliable system even at an Exascale.
  - The biggest challenge is that architectures must be willing to pay the price of hardware detection.

Water cooling

- Water cooling will be “standard” in the future.
  - Not because racks get hotter…. Because it results in better overall power efficiency and cost efficiency.
- Need standards to allow for facilities to choose an optimal, machine independent cooling infrastructure.
  - Cooling building water for rack cooling is a waster of energy.

Case 1. Air Cooling



Total "overhead" cooling power

$P_{cool} = P_1 + P_2 + P_3 + P_4$

Key:

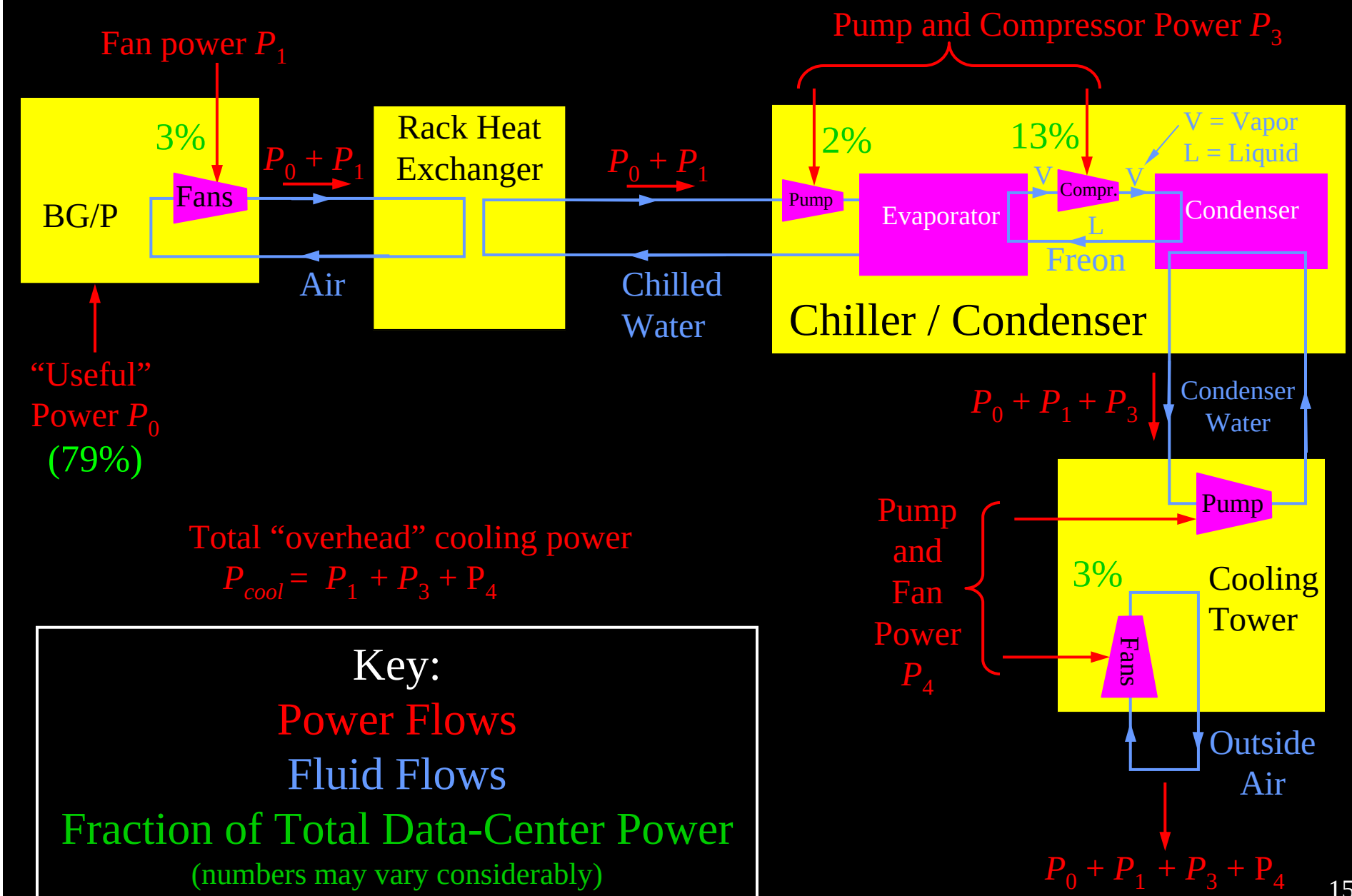
Power Flows

Fluid Flows

Fraction of Total Data-Center Power

(numbers may vary considerably)

Case 2. Hydro-Air Cooling Using Chilled Water



Total "overhead" cooling power

$P_{cool} = P_1 + P_3 + P_4$

Key:

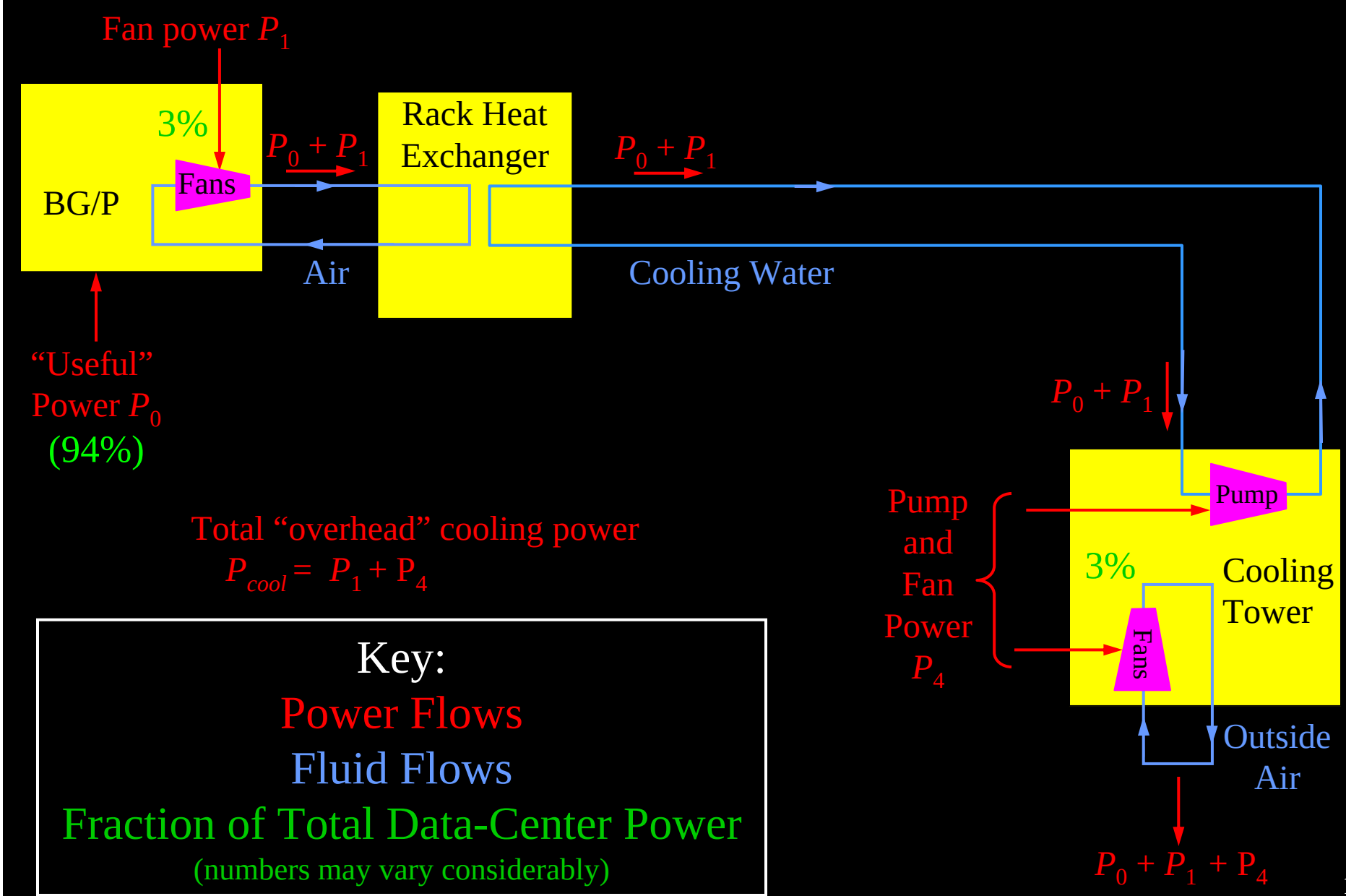
Power Flows

Fluid Flows

Fraction of Total Data-Center Power

(numbers may vary considerably)

Case 3. Hydro-Air Cooling Using Ambient-Temperature Water



Total "overhead" cooling power

$P_{cool} = P_1 + P_4$

Key:

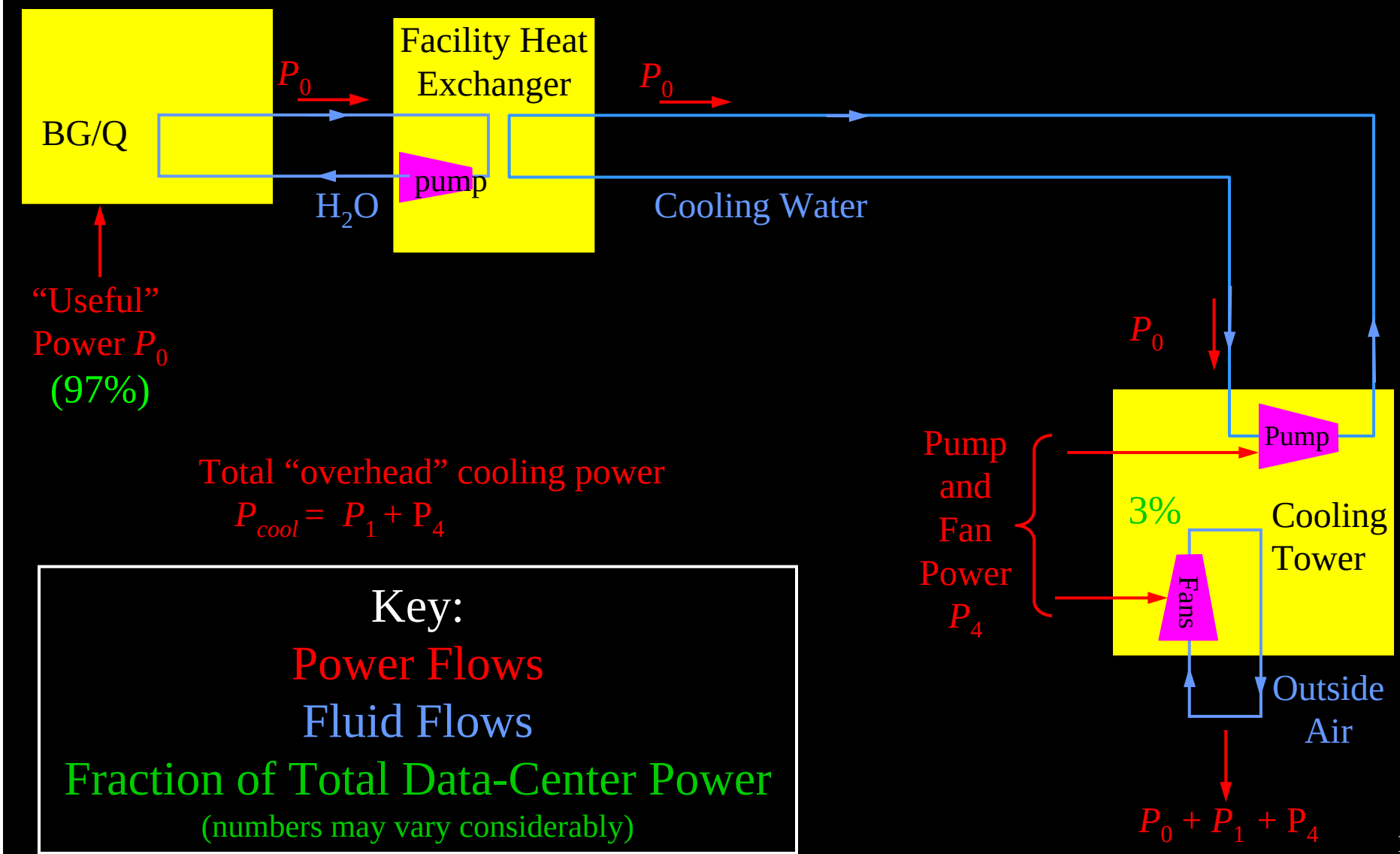
Power Flows

Fluid Flows

Fraction of Total Data-Center Power

(numbers may vary considerably)

Case 4. Indirect H_2O Cooling Using Ambient-Temperature Water



Total "overhead" cooling power

$P_{cool} = P_1 + P_4$

Key:

Power Flows

Fluid Flows

Fraction of Total Data-Center Power

(numbers may vary considerably)

Summary

- The challenges to building an exascale system with reasonable power are numerous.
  - We need to work aggressively with memory manufacturers to find acceptable power solutions.
  - System level Perf/Watt must be elevated to the highest level constraint.
  - There are a multitude of system elements that each have their own scaling behavior. All need to be addressed.
- The system level power variation will increase.
  - Optimizing for real work loads will mean moving past some industry business as usual trends. (Worst case design practices)
  - Power optimized design techniques, compilers and tools need to all be dramatically improved.
- We have seen about a 2.5x system level power efficiency improvement over the last 3 years. We need about 100x improvement over the next 10 years to get to a 20 MW Exaflop system.