TRUE NORTH PROGNOSTICS

EXECUTIVE OVERVIEW

Rethink Cooling to Boost Efficiency

- Think of your data center as a manufacturing facility where there are several inputs (raw materials, pre-fabricated parts, people, etc.) that are combined/processed to create ‘widgets’. In a data center context, what is being produced are completed computer ‘jobs’ and/or running on-line computer programs (widgets) where the objective is to produce the greatest number of widgets at the lowest cost in the shortest amount of time.
- The PUE™ ratio was established by the industry in the early 2000s to measure data center efficiency. Saving energy by reducing PUE™ ratio was universally a good objective when computer chips ran at the same operating frequency, regardless of ambient temperature, when the chips did not throttle, the fan speeds were constant.
- The technology has evolved since the inception of the PUE™ metric. Chips run at much higher operating frequencies today, creating more heat and now need to be throttled (even significantly up to 80%) to avoid damage. When the ambient temperature is increased (to save energy and reduce the PUE™ ratio), it causes the fan speeds to increase which in turn creates increased vibrations that slow down the storage I/Os, causing all the programs to run much longer and consume more energy.
- As a result, this “throttling of chips” **has introduced a change the relationship of how IT components consume power**, making the PUE™ ratio less an indicator of efficiency today.

Rethink Cooling to Boost Efficiency

- Over the years, many data centers have tried to cut costs by raising ambient temperatures to reduce cooling expenses and thereby lower their PUE™ ratio. While increasing temperatures in the past (before chips began to throttle) may have saved energy from reduced HVAC cooling costs. However, there's a smarter and more efficient way to achieve greater savings (lower overall energy costs & improved performance) with today's technology.
- The Optimized Power Management (OPM™) solution does not address the PUE™ ratio because it does not measure the entirety of the work being performed by the data center. In other words, PUE™ does not account for the entire process of the widget production. *(for more discussion on PUE™, see the attached Appendix).*
- Using OPM™, energy use is continuously optimized across the entire data center using real-time telemetry data (e.g., distributed internal currents, voltages, temperatures, fan speeds, etc.) to balance the combined energy of the chips, fans, and I/O against CRAC energy by adjusting the ambient temperature of the data center dynamically.

Rethink Cooling to Boost Efficiency

Savings:

- Up to 30% (or more) reduction of total electricity costs → O&M savings
- Faster job completion → Improved equipment utilization
- Fewer new servers and additional hardware needed → Capital savings

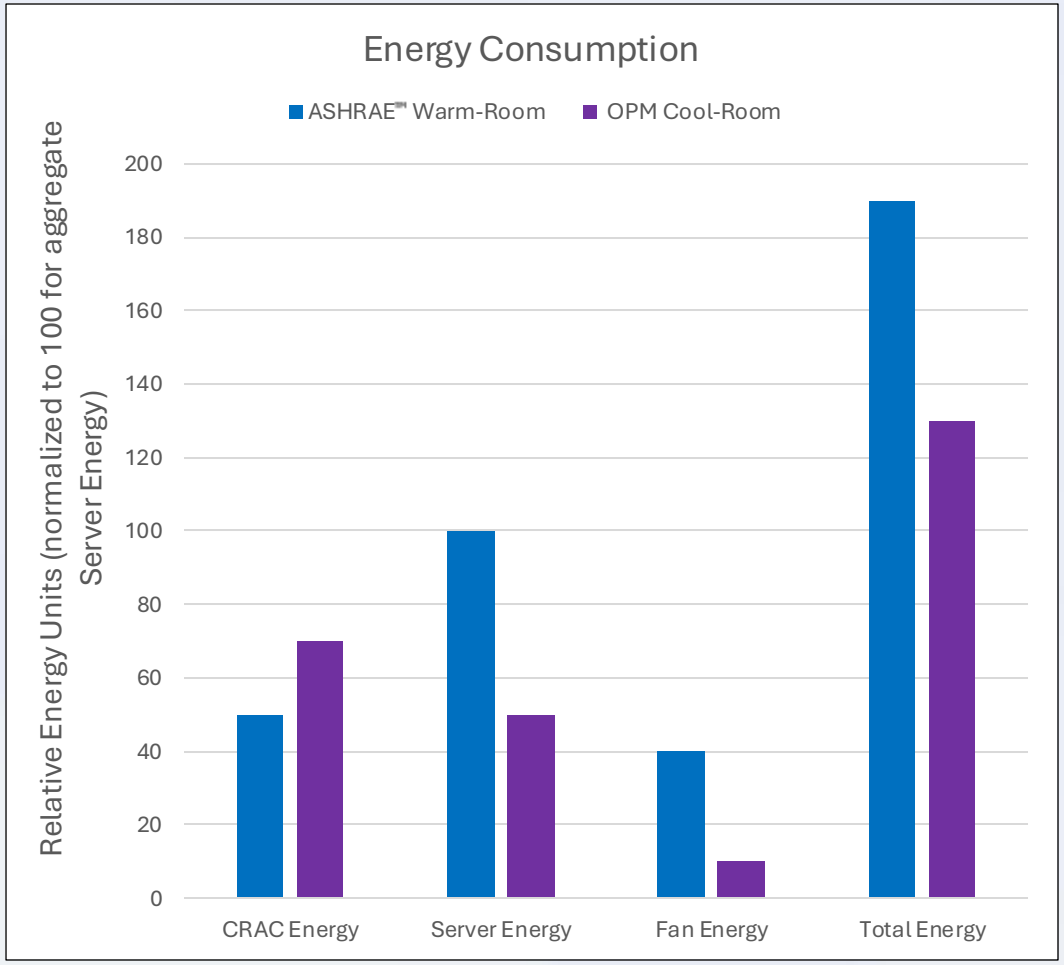
OPTIMIZED POWER MANAGEMENT ™

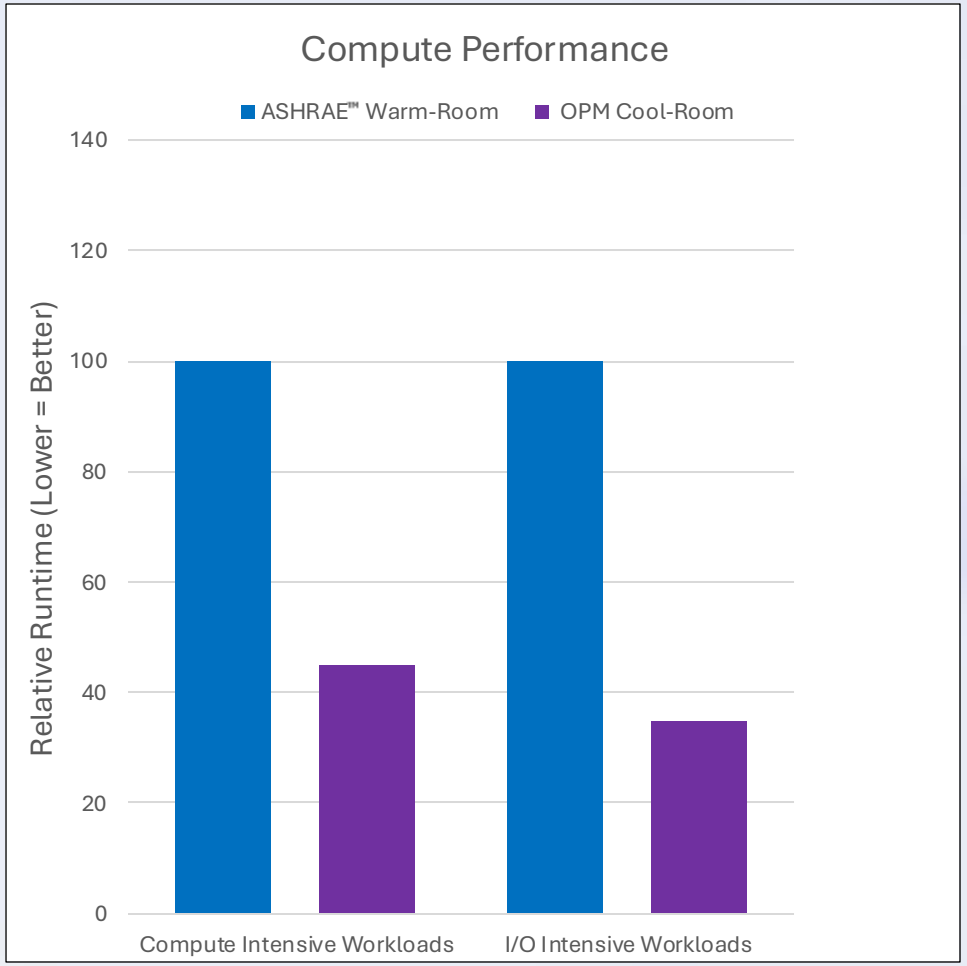
OPM™ OPTIMIZES BOTH PERFORMANCE AND ENERGY EFFICIENCY

Key Takeaway:

Cooling data centers intelligently with OPM™ saves far more energy than warm-room approaches.

By preventing throttling, Improving I/O, and slashing fan power, OPM™ optimizes both performance and efficiency.





ASHRAE™ - American Society of Heating, Refrigerating and Air-Conditioning Engineers

APPENDIX A

THE PROBLEM

- 2008 ASHRAE™ warm-room guidelines suggest decreasing HVAC energy costs^
- Hotter servers begin to throttle CPUs/GPUs → often doubling or more workload runtimes
- Aggregate fan-motor power rises *cubically* with fan speeds → steep energy penalties
- Increased fan speeds create more vibrations that slow I/O throughput → substantially increasing workload runtimes and wasting energy
- **Net result**: More Megawatt-Hours are consumed across the data center, not less

^ https://www.energystar.gov/products/data_center_equipment/5-simple-ways-avoid-energy-waste-your-data-center/raise-temperature?utm

TRUE NORTH PROGNOSTICS

HISTORY OF PUE™ – POWER USAGE EFFECTIVENESS

$$\text{PUE}^{™} = \frac{\text{Total Facility Energy}}{\text{IT Equipment Energy}}$$

- In 2008, the Green Grid Consortium™ (Intel™, AMD™, HP™, Microsoft™, etc.) introduced Power Usage Effectiveness (PUE™). Defined as the ratio of Total Facility Energy to IT Equipment Energy where the ideal value is as close to 1.0 as possible
- Hyperscale operators like Google™ and Microsoft™ began publishing PUE™ numbers. The industry average was 2.0–2.5, while Google reported ~1.2 {DeepMind™}. PUE™ became a benchmark in sustainability reports and a marketing tool, and standards bodies such as ASHRAE and the EU Code of Conduct for Data Centres began adopting PUE™
- Hyperscale leaders have achieved fleet-wide PUE™ averages around 1.10–1.12
- **HOWEVER**: PUE™ does not measure the data center work (compute transactions and cumulative I/O) being performed, i.e. the work done per energy consumed, or the “Work per Watt™”
- Metering the cumulative compute-and-I/O work being done, demonstrates (and is proved quantitatively for any modern air-cooled servers with any CPUs and/or GPUs), that warming up the data center (to lower the PUE™) wastes more energy in the IT systems than is saved in the cooling systems
- PUE remains a key metric for assessing infrastructure efficiency — but it doesn’t capture computational productivity. Modern frameworks such as ITWC, ITUE, and “Useful Work per Watt” complement PUE by providing a fuller picture of overall efficiency (+ see page 11)

TRUE NORTH PROGNOSTICS

HISTORY OF PUE™ AND WHY IT'S NOT A MEANINGFUL MEASURE TODAY

Reducing PUE™ was universally a good objective with computers and storage from 20 years ago and earlier. Before the mid 2000s, all computer chips ran at the same operating frequency regardless of ambient temperature and I/O rates for spinning disk drives were unaffected by fan vibrations inside the servers and storage arrays. Moreover, almost all server fans were constant speed fans, and the fan motor energy was a small portion of the energy budget for servers and storage

Today, with modern computers and storage, reducing PUE™ by raising ambient temperatures is extremely counter productive because:

- All CPUs and GPUs have dynamic frequency scaling, and the "throttling" goes up exponentially with chip temperatures
- All spinning disk drives have become hyper-sensitive to low-levels of vibrations. As ambient temperatures warm up in the data center, server and storage fans increase their RPMs, raising rack vibration levels, substantially reducing I/O rates
- Thanks to Moore's Law, fan motors (and the numbers of fans) have gone up, and for all fans, the energy for the fan motors goes up with the cubic power of the fan RPMs. Raising the data center ambient temperature significantly increases the "energy wastage" from the fan motors … and all of the heat dissipated from the fan motors incurs a "double penalty" (it adds to the heat that must be removed by the CRACs)
- Most significantly (and not reflected in the PUE™ metrics): when the compute performance for the IT systems is cut in half [from frequency throttling in the chips and I/O degradation in the disk drives], it means that for fixed customer workloads, the completion times double, burning 2x the energy for everything in the servers [memory, power supply units, PCIE I/O cards, ASICs, fan motors, as well as the CPUs and HDDs]

OPTIMIZED POWER MANAGEMENT: RETHINK COOLING TO BOOST EFFICIENCY

Work per Watt™

- By cooling the data center, not warming, the overall data center energy is reduced and all compute and I/O performance for the IT assets are significantly increased
  - With no hardware modifications anywhere in the data center, OPM's™ identifies the optimum ambient temperature that substantially boosts both compute and I/O performance, shortens workload completion times, and significantly reduces the total energy requirements across the data center
  - Moreover, the organizations can reduce their buying of new hardware for the data center to meet business "capacity planning" criteria
- OPM™ saves tons of carbon, but at the same time also saves megawatts which translates into significant monetary savings. It also significantly boosts both compute performance and I/O performance, resulting in substantially greater ROI for the capital investments in the data center equipment
- Today, the PUE™ measurement has become counter-productive when chips started throttling and when storage became hyper-sensitive to vibration levels twenty years ago. New, more modern measurements include meaningful work per watt consumed

CRITICISM OF THE PUE™ METRIC

In technical forums, academic literature, and industry events, you'll find frequent criticism or frustration about PUE's limitations, especially when it's presented as a definitive efficiency metric.

Some examples:

- Academic papers and studies often describe PUE as a "failed metric" or argue that it's obsolete in modern high-density, AI, or cloud data centers. (*fluix.ai*)
- Industry analysts and commentators have called PUE "only half the story" or "not the whole picture." (*Data Centre Magazine*)
- New or extended metrics are frequently proposed because many consider PUE inadequate; critics say relying on it as *the* standard is outdated.
- So, while the general public rarely mocks PUE, within the data center engineering community it's often questioned, challenged, and regarded as insufficient on its own.