© Josip Loncaric

Towards Balancing Power, Performance, and Reliability in System Design

Josip Loncaric, HPC-DO, LANL

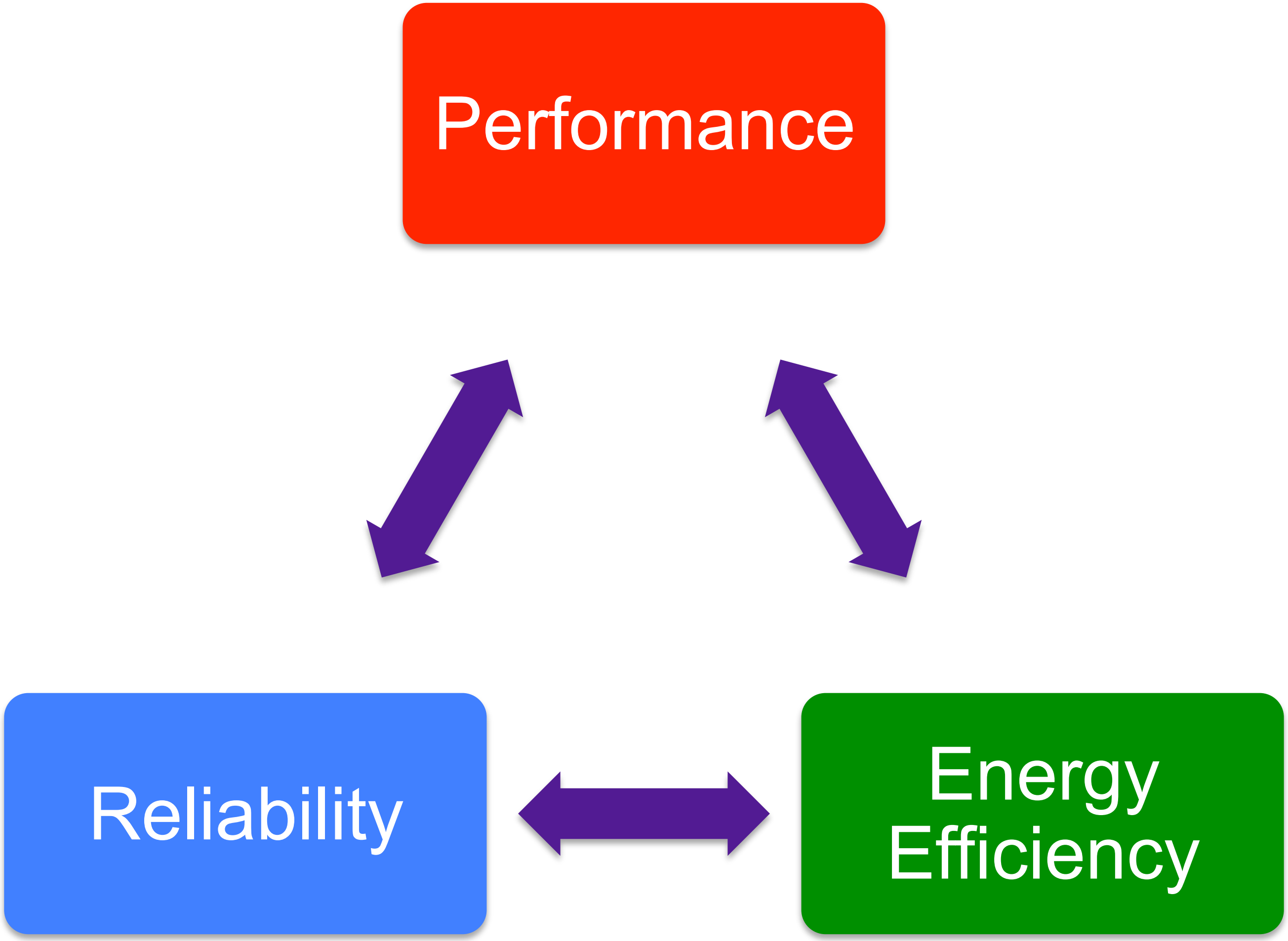
2013 Salishan Conference on High Speed Computing

LA-UR-13-22649

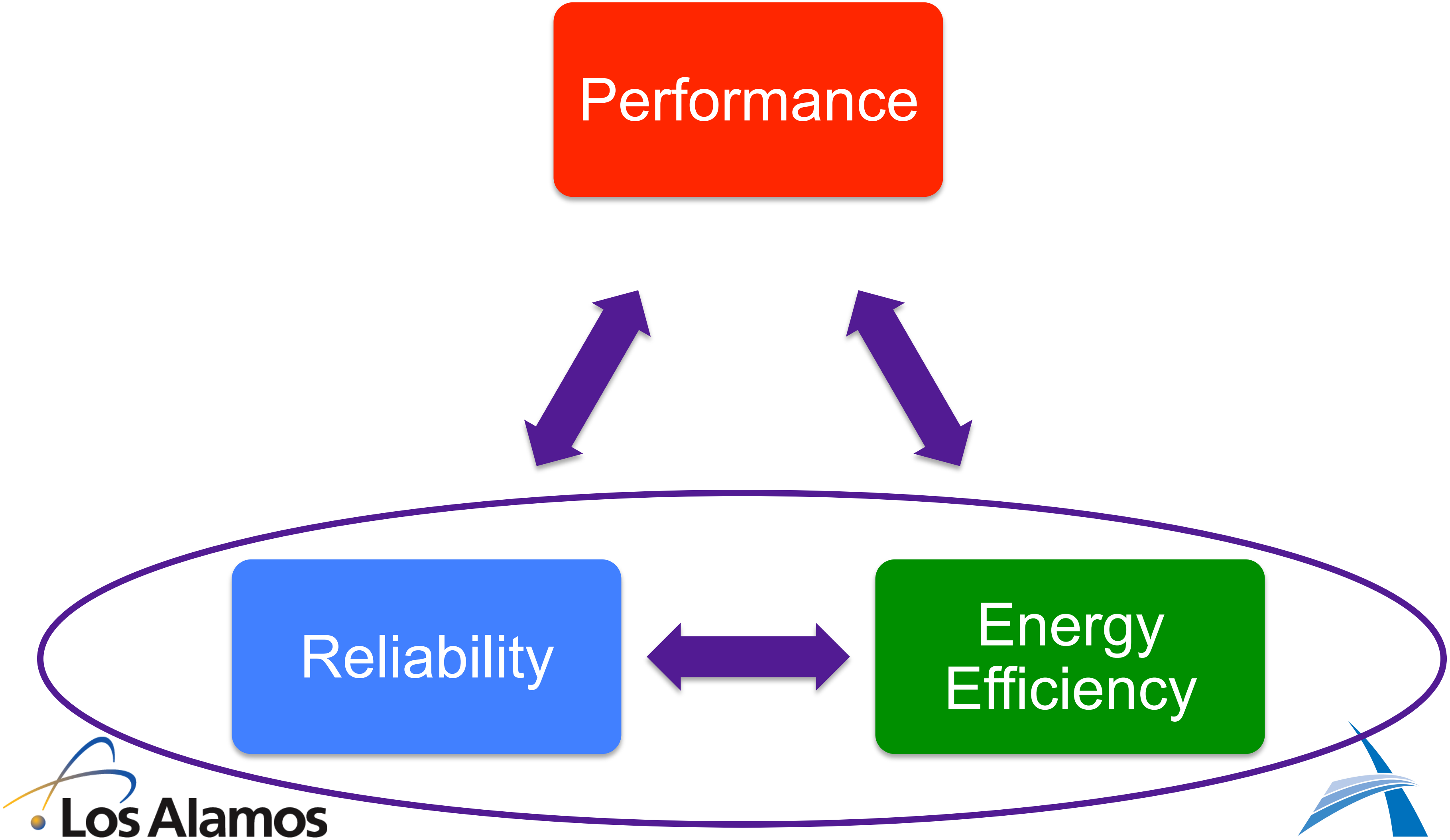
Outline

- Performance / efficiency / reliability
- Getting correct results drives computing
  - Nobody cares how fast an incorrect final result was
- Resilience implications
  - 2-level checkpoint/restart using burst buffer needed at many-PF/s scales
- Power efficiency challenges
  - Complexity of advanced power management
  - Analog environment impacts on reliability of digital logic
  - Unintended consequences & power grid impacts
  - Electricity cost is not simply proportional to energy used
- Conclusions
  - Design to optimize reliable progress under TCO and power constraints

Three Objectives Are in Tension: Take Any Two

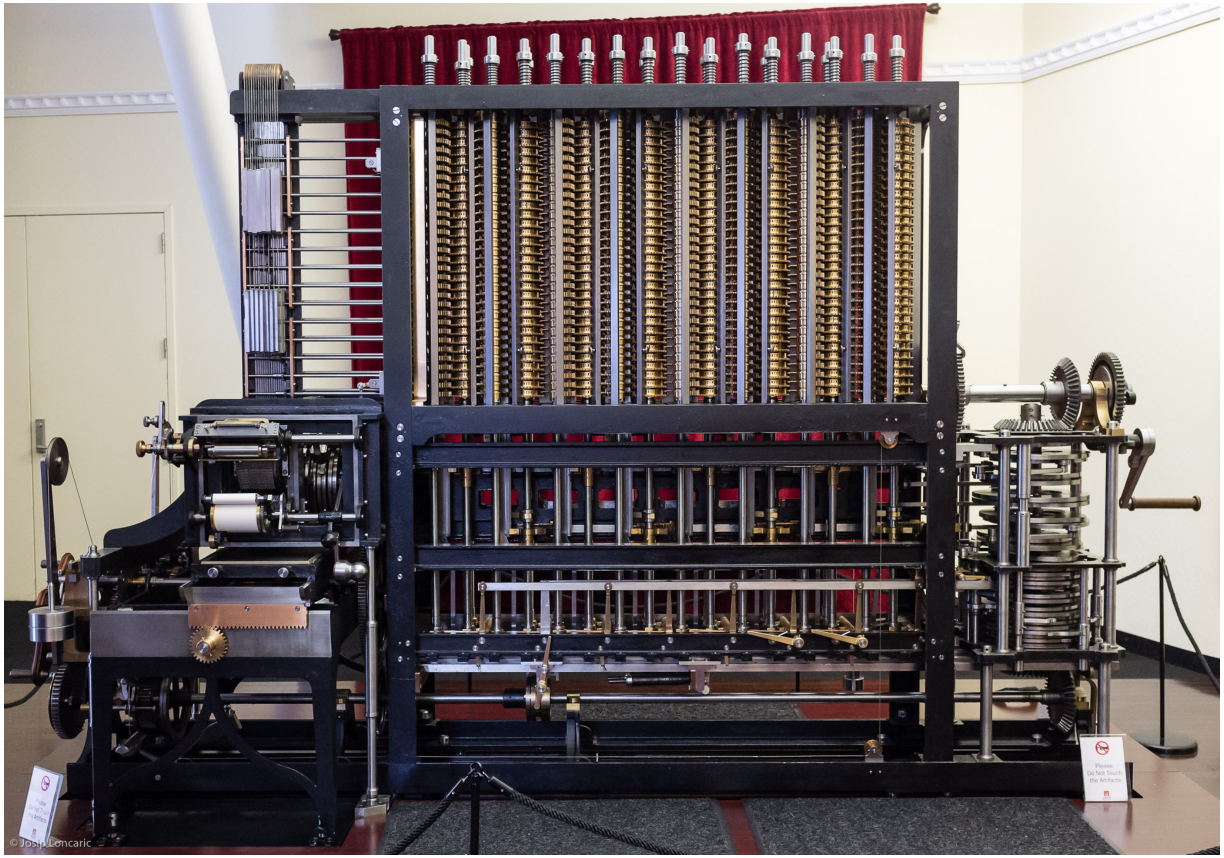


Reliability & Energy Efficiency vs. Performance



Correctness Created Computers

Babbage's early 1812 vision was the "unerring certainty" of machinery

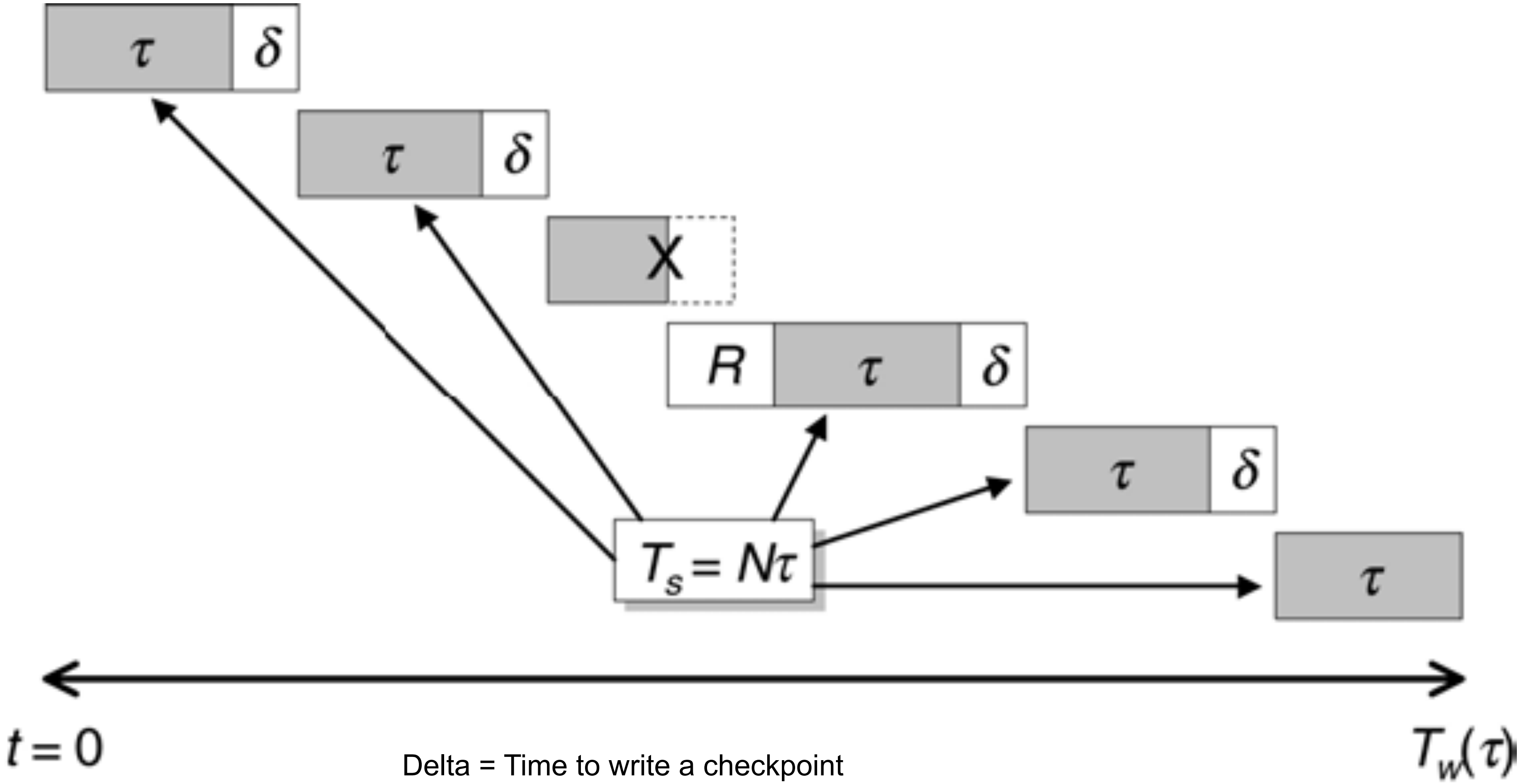


Correctness Demands Digital Computers

Project Whirlwind tested correctness of a 5-bit digital multiplier in 1948

Correctness Extended By Defensive Measures

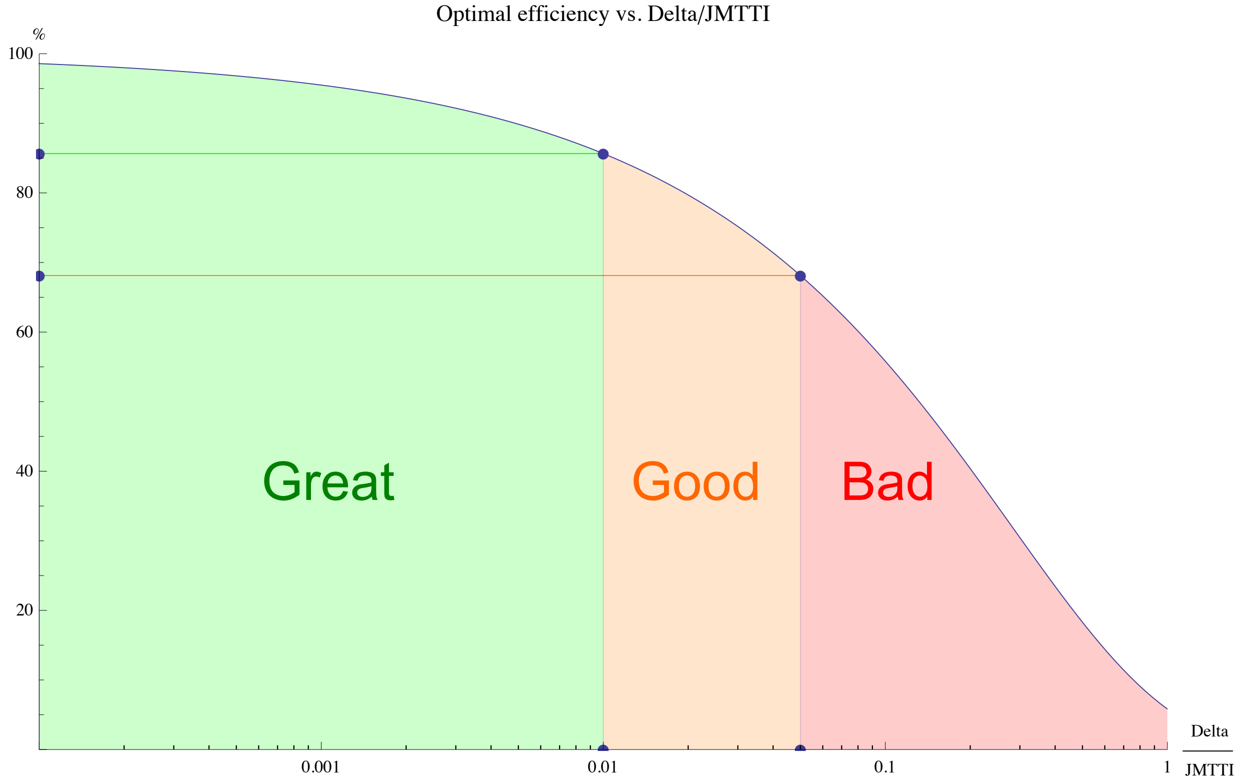
If computation fails, restart from checkpoint and try again



Delta = Time to write a checkpoint
R = Time to restart from a checkpoint (usually ~ Delta)
Tau = Compute time, including rework after restart

Progress Efficiency With Checkpoint / Restart

Exponentially distributed faults + optimal C/R = universal[1] efficiency curve

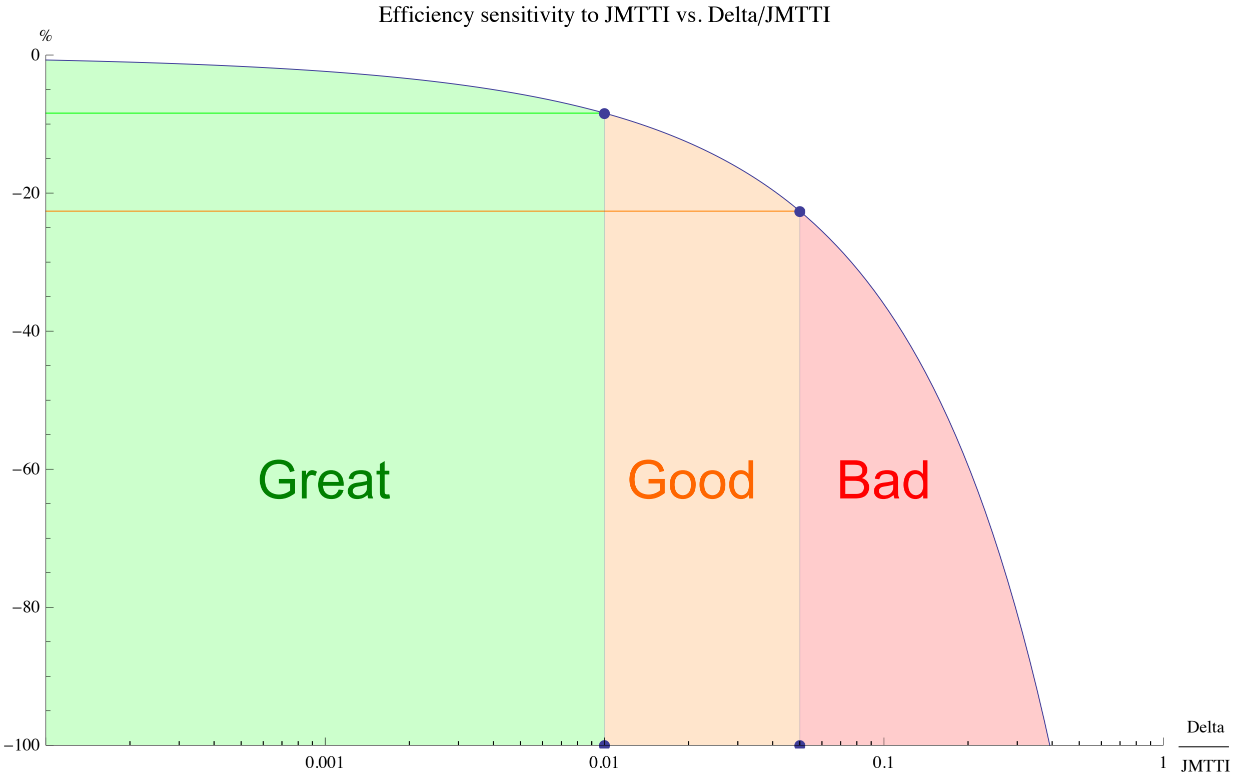


Checkpoint / restart is still the only commonly used resilience technique

[1] Simplifying approximation: Restart and checkpoint times equal

Sensitivity of Efficiency to JMTTI Decrease

Relative decrease of efficiency per fraction decrease in JMTTI at fixed Delta



JMTTI decrease when Delta/JMTTI >> 0.01 has a large relative impact on progress efficiency

Non-dimensional sensitivity of f(x) = (df/f)/(dx/x) = (df/dx)*(x/f)

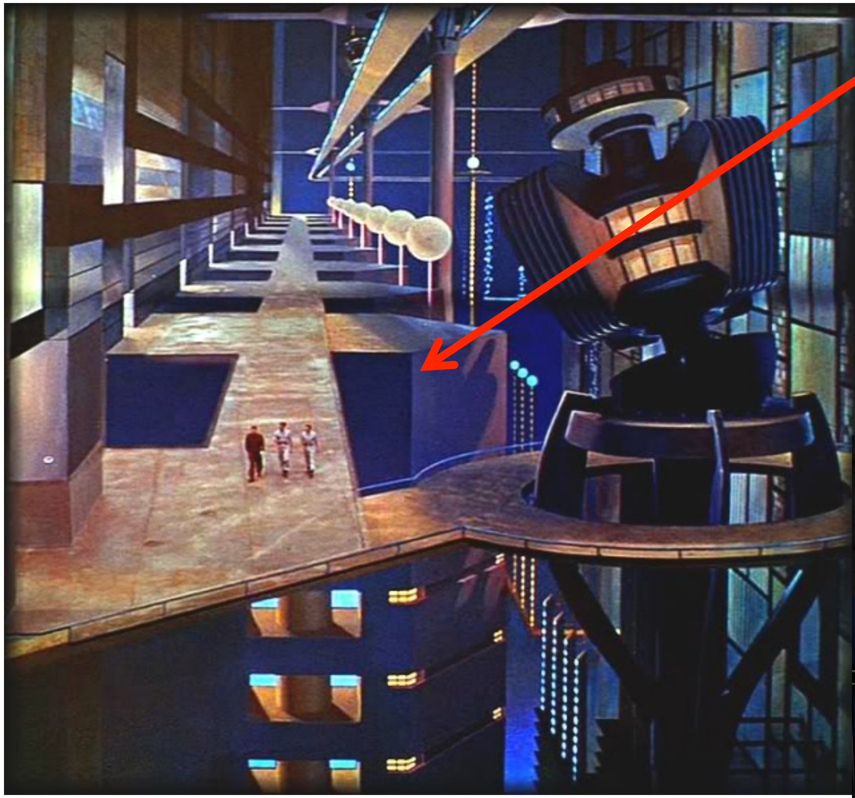
Soon: Efficient Progress Requires Burst Buffer

- Same efficiency at same Delta/JMTTI
  - JMTTI scales as 1/N
  - Memory footprint scales as N
  - Checkpoint N times more memory in 1/N the time

⇒Checkpoint bandwidth scales as N^2

- Disks priced for capacity, not bandwidth
  - N^2 more disk spindles? Cost prohibitive soon.
  - Disk capacity ~30x memory typically suffices, but lacks sufficient bandwidth
- Burst buffer concept
  - Checkpoint to solid state devices (1st level), less often to disk (2nd level)
  - Expands storage hierarchy to get more affordable bandwidth
  - Burst buffer ~3x memory @ ~10x disk bandwidth

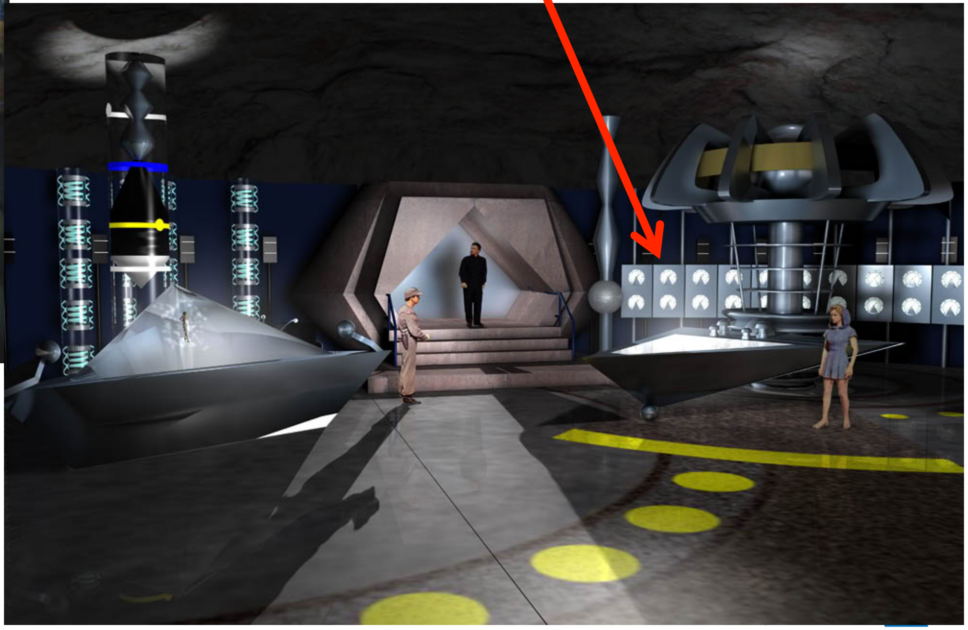
Great Machines Have Great Requirements



Forbidden Planet, 1956

Planetary scale machine

Exponential power demand



Extrapolating HPC Performance & Power Trends

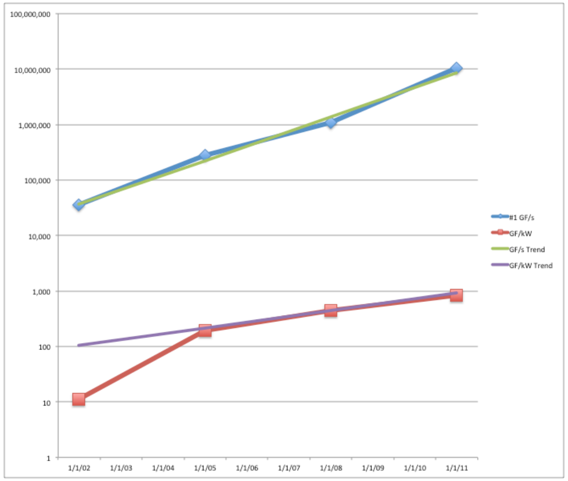
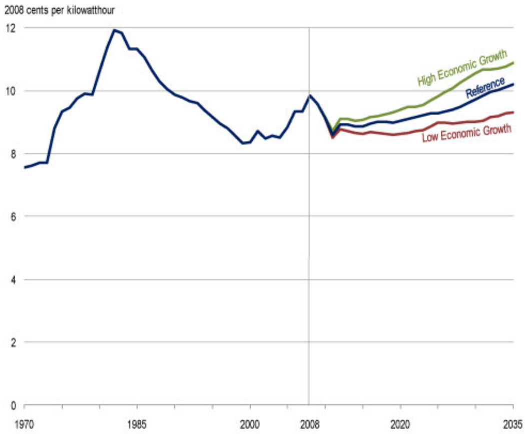
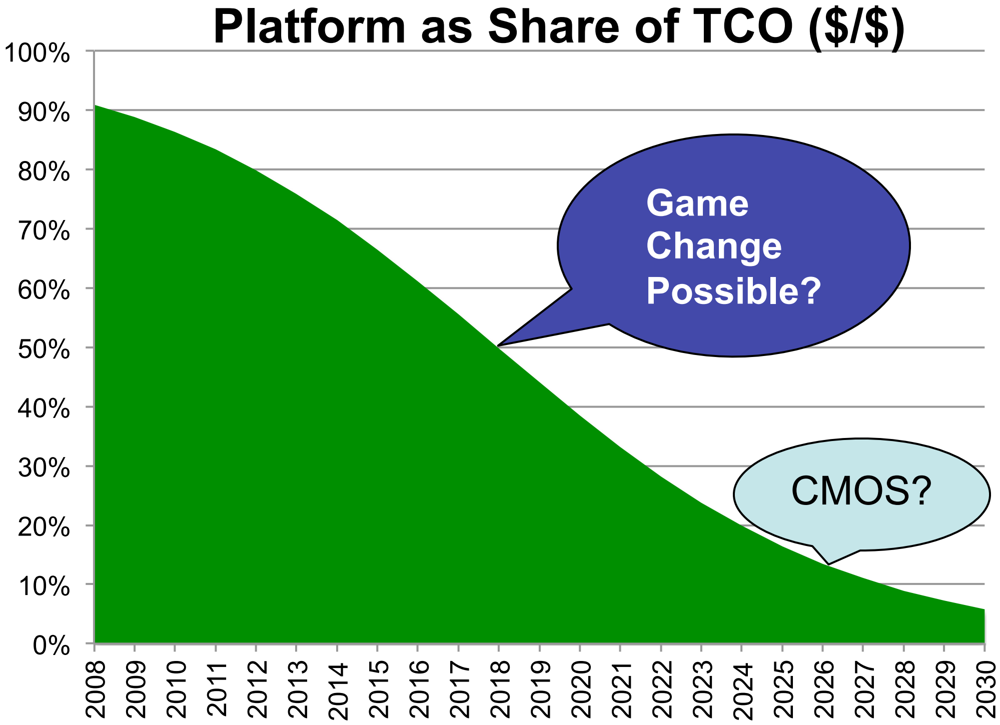


Figure 60. Average annual U.S. retail electricity prices in three cases, 1970-2035





Sources:
(1) Top500 Nov. 2011 list
(2) U.S. Energy Information Administration, Annual Energy Outlook 2010

Energy Efficiency Through Power Management

- Active power ~ C×V^2×f
  - Leakage power is additional
- Capacitance ~ hp
  - Half-pitch scales linearly, at half the pace of Moore's law
- Power efficiency doubles in ~3 years
  - Robert Dennard scaling for CMOS (1974) underlies Moore's law
  - Lots of factors in play, but that's the approximate trend line
- But, what if activity were reduced?
  - Opportunities for reduction in idle power
  - Power management proliferates
  - More power domains, more power states, more transitions, more heuristics
  - Objective: Use power only when needed, and only as much as needed

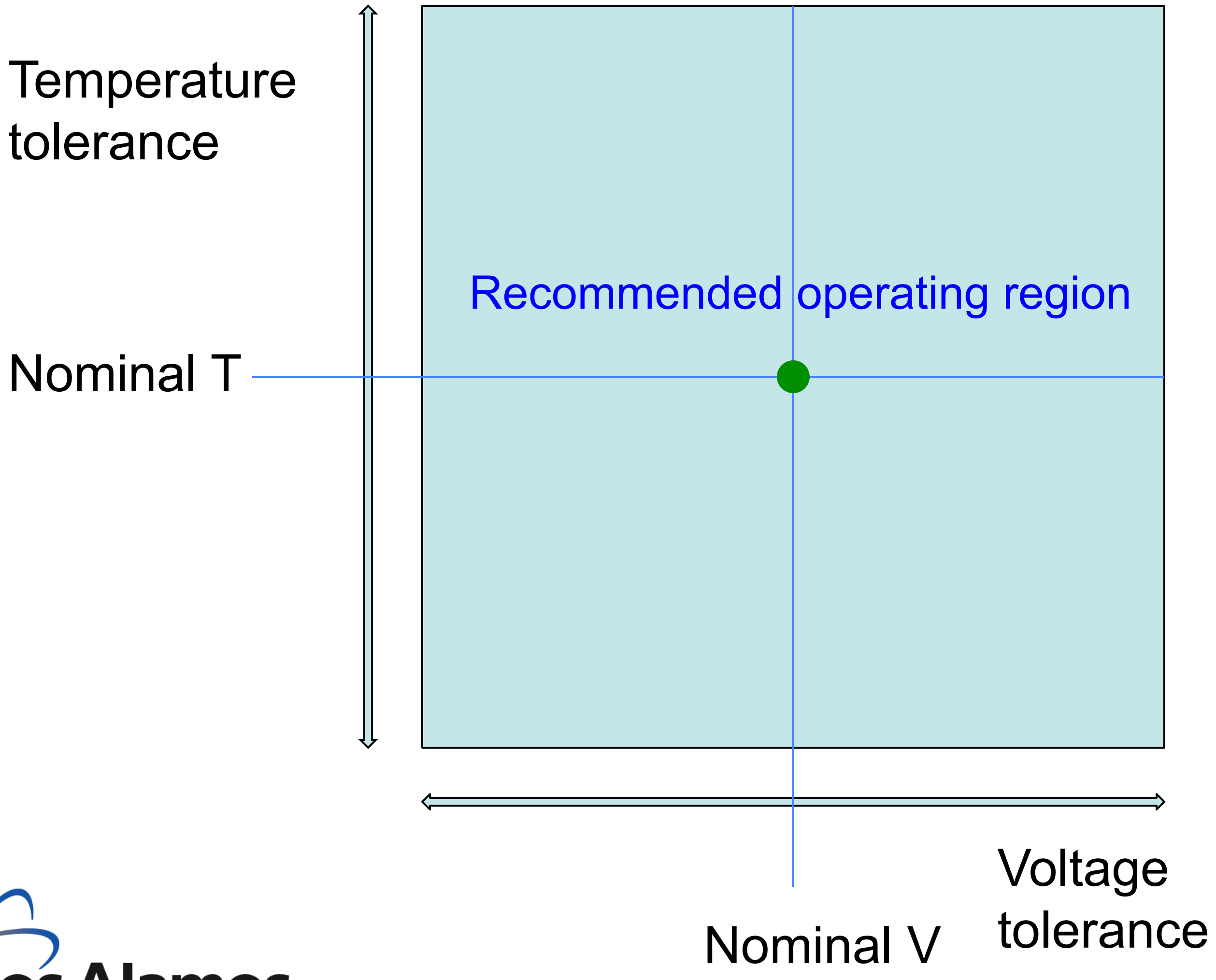
Logic Drowning In a Sea Of Voltage Regulators

- Multiple power domains per chip
  - Typically with power gating per core
  - More power planes for PCIe, memory, etc.
- Monitoring temperature, power
- Accepting system power management requests and limits
- With dynamic control of power states
  - Voltage, frequency optimized under power and thermal constraints
- Complexity:
  - 5-10 power domains today, may grow to 20-50 in the future

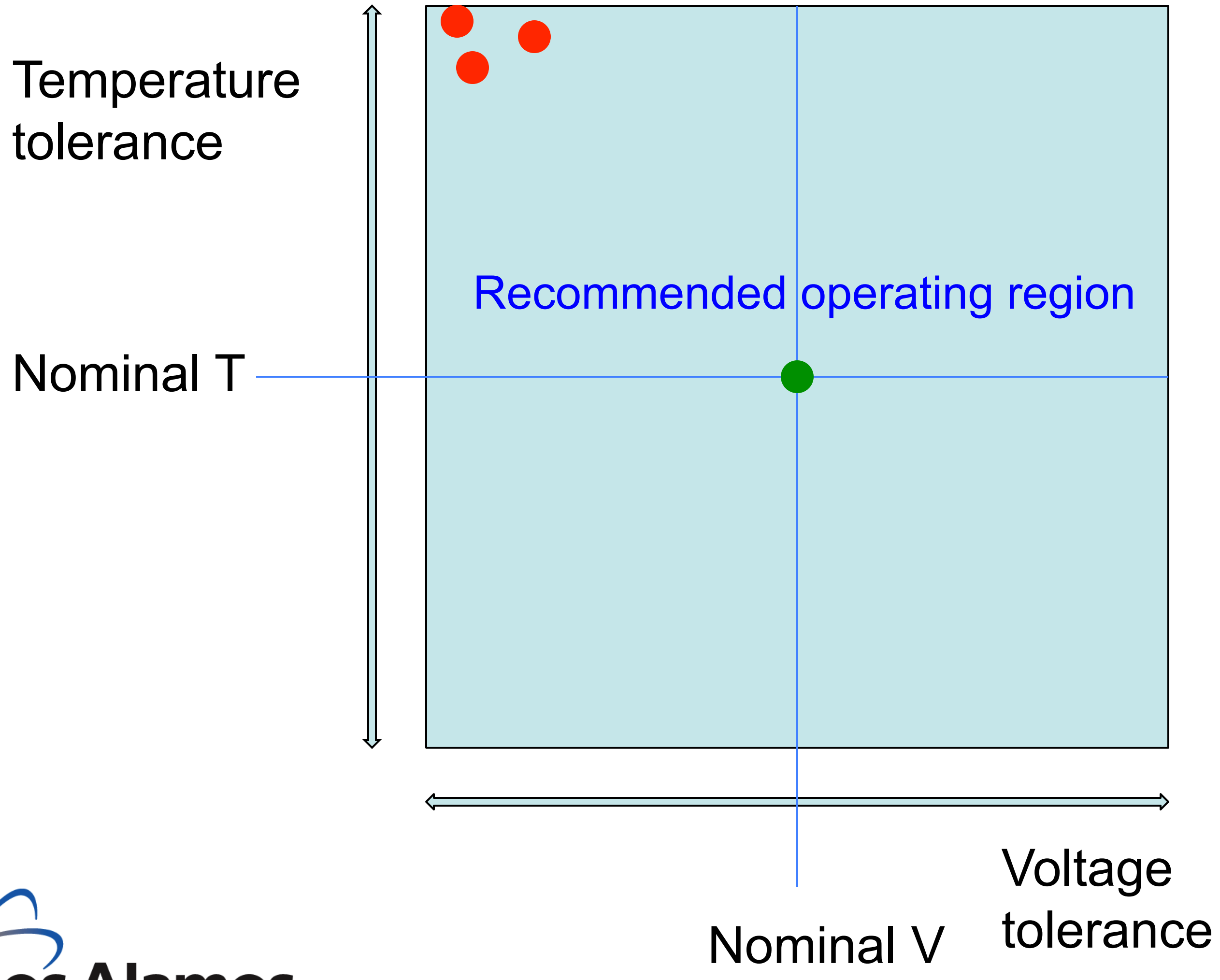
Voltage Regulators and Reliability

- Most often *replaced* HPC parts, in order:
  1. Memory
  2. Anything with a voltage regulator: Power supply, motherboard, …
  3. Network
  4. CPU
  5. […]
- More voltage regulators ➔ less reliability
  - Consolidate voltage regulation in fewer, more reliable parts?
- Daunting complexity of verifying systems
  - Digital electronics live in analog world of voltages, temperatures, etc.
  - Precision demands are growing, coordination demands are growing, …
  - Operating conditions have many corner cases, vary dynamically
  - Examples follow

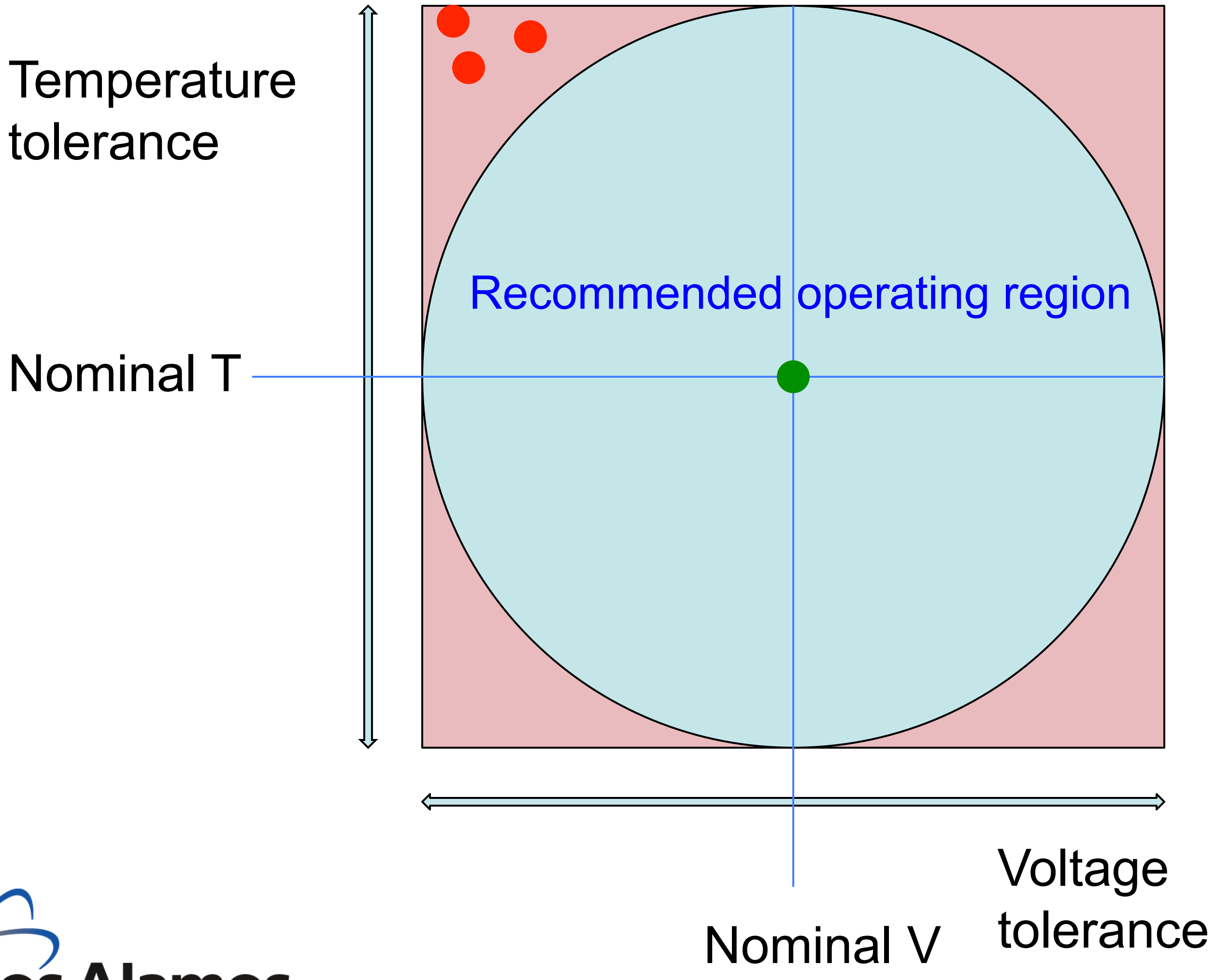
Simple Case: Two Analog Variables



Unreliable Operation Seen in Corners



Redefine Tolerance to Exclude Corners



In 2D:

Slight reduction in operating region by factor: π/4

Looming Disaster: N Analog Variables

- Modern CPUs have multiple voltages
  - Each voltage (or local temperature) has its tolerance
  - N-dimensional cube with inscribed N-dimensional sphere
- Volume ratio:

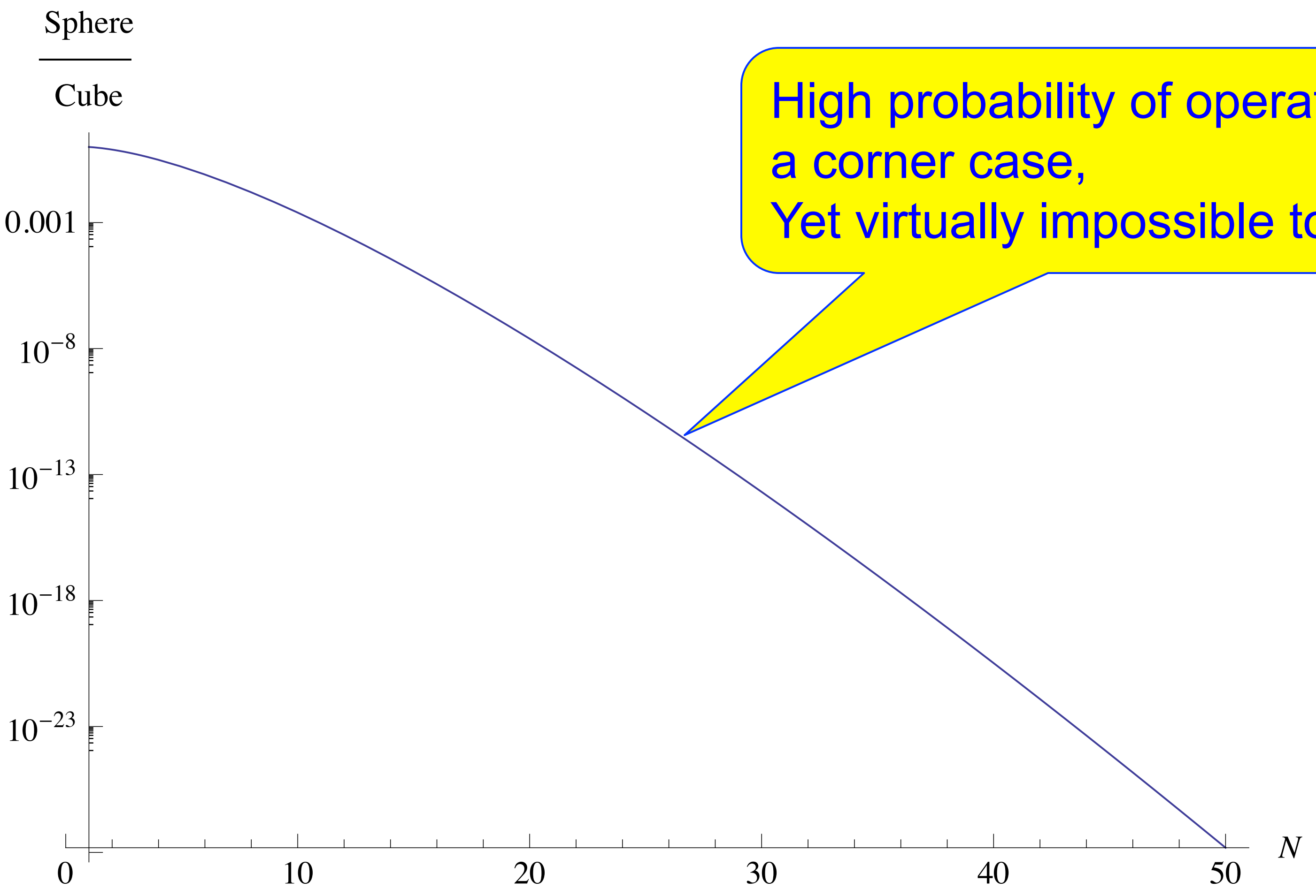
$$\frac{Sphere}{Cube} = \frac{\left(\frac{\sqrt{\pi}}{2}\right)^N}{\Gamma\left(1+\frac{N}{2}\right)}$$

- Nearly ALL volume is in the corners:

N	1	2	3	4	5	6
Sphere	100%	78%	52%	31%	16%	8%

TODAY!

Corner Cases Dominate as N Grows



High probability of operating in a corner case,
Yet virtually impossible to test

Can It Get Worse? Yes!

- Analog variables are not static
  - Need detection & prompt reporting of *analog* variables out of tolerance, under transient operating conditions
- CPUs specify both DC and AC tolerance
  - Power demand fluctuates at CPU rates
- Voltages vary at DC-MHz frequencies
  - What is the *worst* case frequency for voltage regulation?
  - We can and sometimes *we do* hit worst cases
  - Bizarre non-repeatable intermittent errors can result
  - Shrinking voltages, tighter tolerances: growing concern
- Need high performance voltage regulation
  - Cost & reliability at 20-50 dynamically changing voltage domains per chip?
  - Guaranteed over 100,000 chips operating for 5 years?
  - Many “eggs in one basket” ➔ **Watch that basket!**
  - Consolidating power management in higher reliability parts should help

Power Efficient HPC Challenge To Power Grid

- Power efficiency gains through:
  - Dennard scaling for CMOS + voltage reductions
  - Power management (DVFS, run to idle) + low idle power
  - Saving many 10's of Watts per chip, low 100's of Watts when idle
  - Nice, but beware of unintended consequences!
- Large scale HPC applications have global synchronization points
  - Synchronous total power demand fluctuations
  - LANL ongoing power monitoring: 1+ MW transients daily
  - Extrapolating to 10+ MW transients soon
  - Can the power grid handle it and deliver good power quality?
  - Initial testing at LANL (Aug. 2012, Dec. 2012)
  - 1+ MW transients seen in a single AC cycle (~15 ms)

Daily 1+ MW Transients Growing to 10's of MW

Large HPC platforms consume large amounts of electrical power

Many HPC applications have global synchronization points

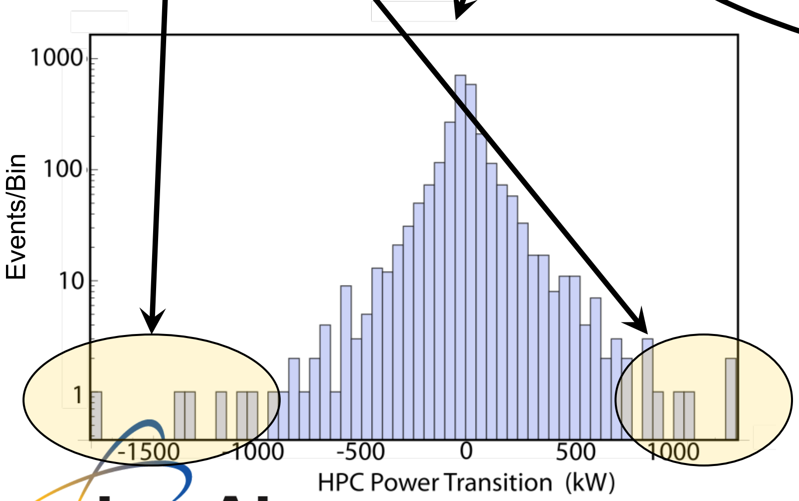
Energy efficiency improved via reduction of CPU idle power

<u>A new class of potentially disruptive grid transients emerging</u>

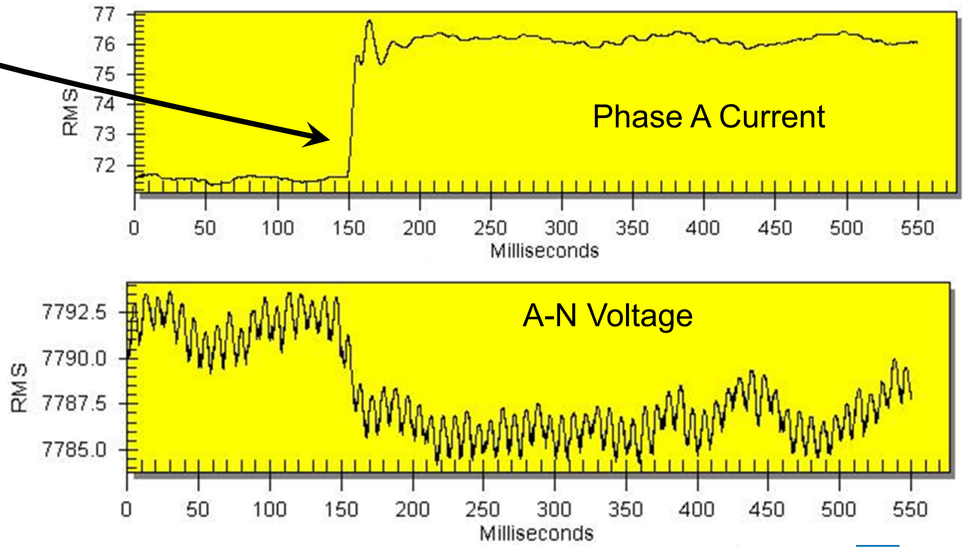
Large—the entire platform (10's MW)

Fast–about one AC cycle (~15 msec)

Current LANL platform experiences ~ "full-machine" transients daily



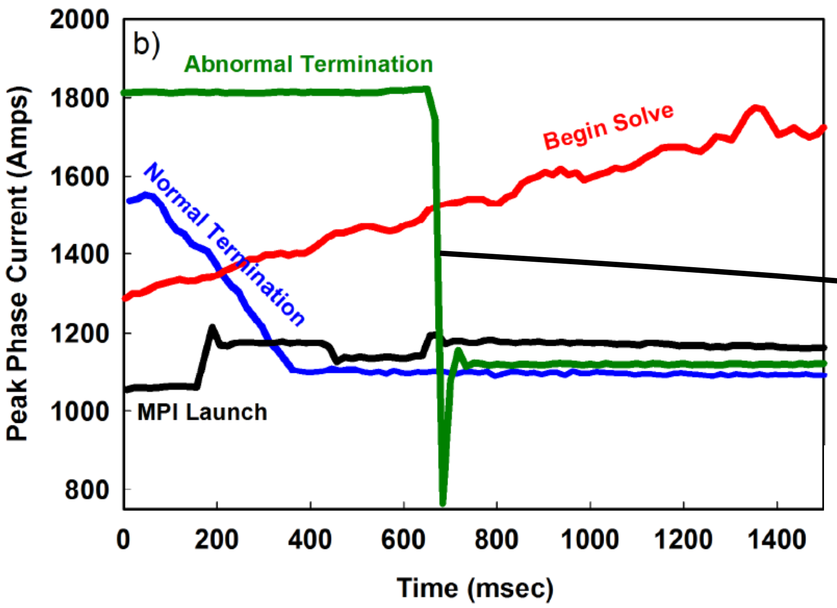
A ~100 kW, single-cycle transition on LANL HPC captured on utility meters at a LANL substation



Los Alamos NATIONAL LABORATORY EST. 1943

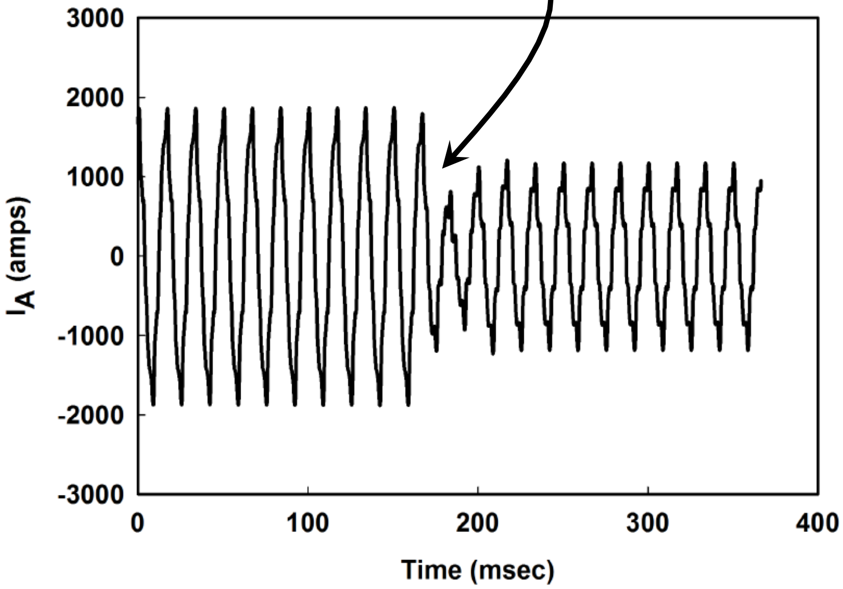
ASC™

Linpack Power Transient Testing at LANL



Projections for LANL platforms:

- Today: Transients not a concern
- 2015: Transients noticeable, still within limits
- >2015:Transients likely need to be mitigated (depending on MW growth)



Joint work with Scott Backhaus, Cornell Wright, and Maura Miller at LANL

Power Efficiency Could Get Expensive

- New problems introduced
  - Power contract pricing: Demand charge + energy charge
    - Demanding giant power fluctuations from the power grid costs serious money
      - Worst case found at LANL: Pay 300+ times more per MWh by pushing demand up briefly at a bad time
  - Fast voltage fluctuations within a large HPC data center will grow
    - Guiding thought behind National Electrical Code: ≤5% voltage sag at load
    - Normal power grid compensation for voltage sags/swells takes ~1 minute or more
  - Combined with external power disturbances, may exceed P/S tolerances
    - Computer P/S normally designed for 10% voltage tolerance >10 sec. duration
    - Data centers require good power quality for reliable operation
  - At 10's of MW: Potential impact on transient stability of the regional power grid ***if*** the grid is already in a stressed state
  - At LANL, we're working towards continued monitoring of power quality and large+fast power transients to understand power requirements and extrapolate to the future
- Preliminary recommendations
  - Local power system requirements should go beyond capacity provisioning to include transient and voltage sag/swell impacts.
  - Consider informing and/or working with ISO/RTO (Transmission operator) to see if frequent real power transients need to be included in their reliability screening, e.g. as an N-1-1 event.

Power Cost Is Not Proportional To Energy

- Power cost = Demand $ + Energy $
  - Demand charge is based on maximum coincidental hourly power reached within a month for the overall power system under the power contract which includes LANL
    - Utilities prefer more predictable power demand, since higher demand requires higher system capacity
    - Managing HPC power use, particularly during the 10am-2pm time slot when peaks typically occur
    - Average hourly peak formula can accommodate short duration power spikes
  - Unexpected power demand fluctuation requires people in the loop, to buy or sell power on the spot market at least one hour in advance
    - Utilities want advance notice of large power changes
    - More advance notice is better and typically results in lower energy cost, looking for >2hr notice
    - Utility may want demand response capability (brief power delivery delays at times)
  - Energy charge is proportional within the power band set by the utility, at variable market rates locked-in the previous day
    - Our power band is ± 2 MW wide
    - Below the power band: Pay for unused energy ($) or sell it (¢) on the spot market, if able
    - Above the power band: Must buy energy at utility’s emergency supply rate ($$$)
  - Daytime energy costs 70% more than nightime energy, on the average
- What we really need is overall *cost* efficiency
  - Efficiency: Total delivered value per total cost

Los Alamos NATIONAL LABORATORY EST. 1943

ASC™

TCO = CapEx + OpEx

- Operational share of TCO is rising:
  - 2008: 90% CapEx, 10% OpEx
  - 2018: 50% CapEx, 50% OpEx
- DVFS costs and benefits:
  - Code-dependent slowdown
  - Various overall workload energy savings estimates, say 10% average saved
  - Is it worth it on the TCO basis?
  - 2008: Not worth it unless slowdown is <1%
  - 2018: May be worth it if slowdown is <10%
- Again, is it worth it?
  - Above, TCO doesn’t include cost of facility, staff, programmatic imperatives
  - Complex management strategies can introduce fragility
  - Need robustness and low performance impact to justify DVFS
  - Possible, if slowdown is off the critical path within the application
  - Advanced power management strategies need to evolve, consult application

Performance / Efficiency / Reliability Conflict

- In a power constrained regime:
  - More performance requires more power efficiency
  - More voltage domains and complex power management approach
  - More complexity leads to less reliability, testing challenges
  - Less reliability leads to more frequent application interrupts
  - Decreased JMTTI reduces progress efficiency due to defensive overheads
  - Decreased progress efficiency reduces *reliable* performance, with large impact when Delta/JMTTI >> 0.01 --- *and wastes more energy on defensive measures!*
- Given TCO, designs should be optimized for delivered *reliable performance*
  - Burst Buffer, reliable electronics, good power quality, monitoring, advanced power management, power contracts, agile adaptivity to external conditions

Conclusions

- Maximizing delivered value per TCO $ requires recognizing couplings between performance, efficiency and reliability
- Impact of power efficiency developments on reliability is unknown but significant
- Until better techniques are available, C/R with checkpointing times ~ 1% of JMTTI avoids high sensitivity to uncertainties
- Burst buffer aims to reach this regime for many-PF/s scale platforms

Abstract

Power is now limiting growth in computing performance, leading to industry-wide push to improve power efficiency of computation beyond classic CMOS scaling. One of the most prominent approaches is the proliferation of power management techniques intended to rapidly reduce power when chips have nothing to do. In practice, this has the potential to reduce system reliability and therefore productivity. We will illustrate some of these operational impacts, and argue that the useful metric for designing balanced systems is total cost of ownership per expected delivered value to end users.