



Breaking New Ground for Scale

Lee Ward



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.





Rumors

- Threads, threads, and more threads
 - With many cores to execute them
 - Always a waiting thread for your computing pleasure
- Infinite memory bandwidth
 - How is not clear, but we assure you it's going to be this way
- Bulk-synchronous checkpoints are a thing of the past
 - As soon as we figure out how to deal with error



Resilience

- We're told the exascale machine will have an MTTI near that of today's
- Do we believe this?
 - It seems we should not
 - This is a bet we cannot afford to lose
- Client is most likely component to fail
 - So we will want a stateful solution because it will be much more fun!



Noise (Jitter)

- It's always been about competition
 - Coupled apps need determinism from the environment
 - So IO needs to stay out of the way
- In the past competition has been about the CPU
- Large core counts can mitigate the historical competition
- The new arena involves memory and network bandwidth
- Jitter concerns won't disappear, only transform



Heterogeneity

- We're told there will be usable amounts of non-volatile memory on-node
- How do you leverage that?
 - Exposed as addressable memory?
 - Persistent?
 - Memory-mapped I/O?
 - Scratch/burst-buffer/staging?



Can we have a new file system, please?

- One that:
 - Is successful at hiding faults at scale
 - Can we “fail in place,” self-organize?
 - Adapt to instead of mandate the environment?
 - Can leverage a heterogeneous mix of media
 - Size, bandwidth, latency, persistence
 - And heterogeneous networks, simultaneously
 - Has a scalable namespace
 - Haven’t we had enough of tree-based yet?



Access Driven

- PLFS and SCR success should be telling us what needs to come
 - PLFS tells us we have improperly constrained the problem
 - SCR tells us we haven't sufficiently considered the opportunities in our machine architectures
- Shouldn't this motivate us to re-evaluate checkpoint and data stream capture solutions
- Can we address, finally, visualization simultaneous with compute on the same file system?



Summary

- Seems we have new opportunities
 - Threads, non-volatile memory on-node
- Old issues, but magnified and transformed
 - RAS, RAS, more RAS
 - Noise
- Opportunities with issues :)
 - Heterogeneous stores
- New opportunities, different (scale)
presentation of old issues **should** motivate
new solutions