Ph.D. Final Defense

Fault Tolerance in Linear Algebraic Methods using Erasure Coded Computations

Ph.D. Candidate | **Xuejiao Kang**

Advisor | Ananth Grama

PURDUE UNIVERSITY

Department of Computer Science

November 26, 2018

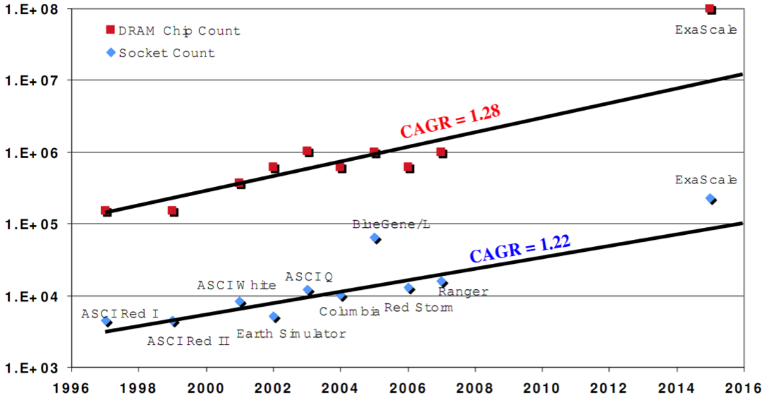
Faults in Parallel and Distributed System

As parallel systems scale to millions of cores, faults become one of the most critical challenges.

As data centers scale to hundreds of thousands of nodes, faults are a prime consideration for distributed computations.

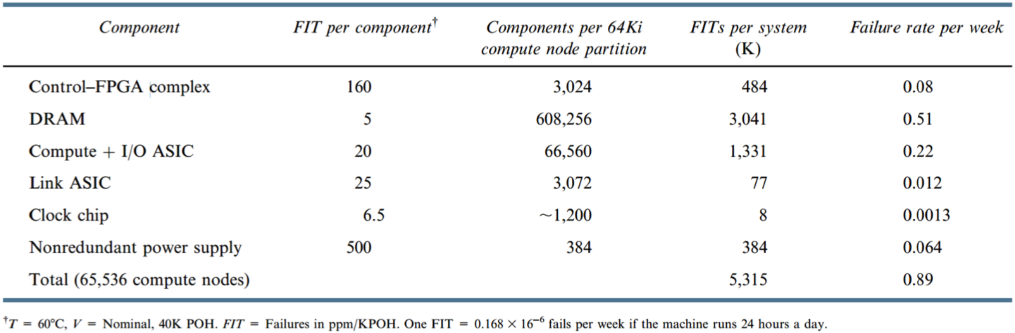
As networks scale from data center to wide area, network faults and partitions constitute a major consideration for wide area distributed computations.

Estimated Chip Counts in Exascale Systems



Source: *DARPA Exascale Technology Study [Kogge et al.]*

BlueGene Failure In Time (FIT) budget

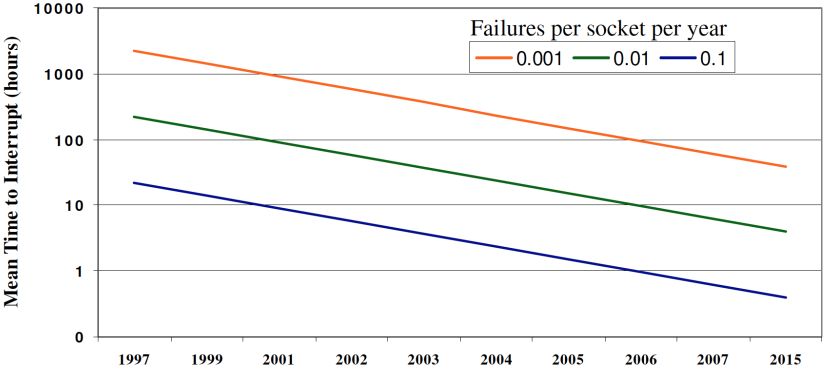


Component	*FIT per component*†	*Components per 64Ki compute node partition*	*FITs per system (K)*	*Failure rate per week*
Control–FPGA complex	160	3,024	484	0.08
DRAM	5	608,256	3,041	0.51
Compute + I/O ASIC	20	66,560	1,331	0.22
Link ASIC	25	3,072	77	0.012
Clock chip	6.5	~1,200	8	0.0013
Nonredundant power supply	500	384	384	0.064
Total (65,536 compute nodes)			5,315	0.89

†T = 60°C, V = Nominal, 40K POH. *FIT* = Failures in ppm/KPOH. One FIT = 0.168×16^{-6} fails per week if the machine runs 24 hours a day.

Source: *P. COTEUS ET AL., IBM J. RES. & DEV. VOL. 49 NO. 2/3*

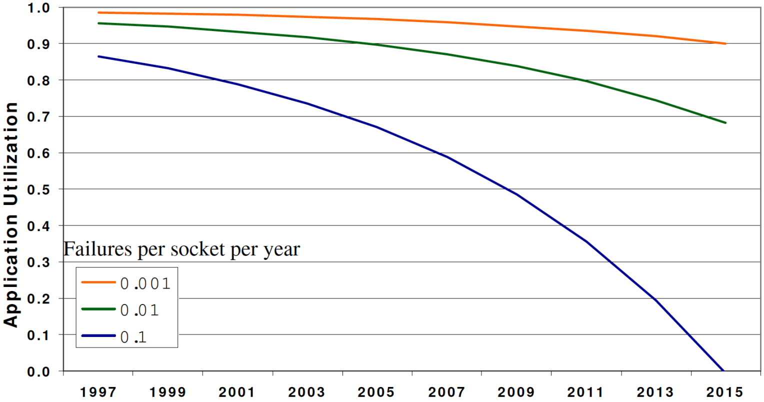
Scaling trends for environmental factors that affect resiliency



Source: *DARPA Exascale Technology Study [Kogge et al.*

Application Utilization for checkpoint overheads

If one socket fails on average every 10 years, application utilization drops to 0 at 220K sockets!



Source: *DARPA Exascale Technology Study [Kogge et al.*