Modeling and Tolerating Heterogeneous Failures in Large Parallel Systems

Eric Heien[1], Derrick Kondo[1], Ana Gainaru[2], Dan LaPine[2], Bill Kramer[2], Franck Cappello[1,2]

[1]INRIA, France [2]UIUC, USA

Context

- Increasing failure rates
  - Est. 1 failure / 30 minutes in new PetaFLOP systems
- Increasing complexity of failures
  - Time and space dynamics
- Increasing cost of failures
  - Est. $12 million lost in Amazon 4-hour outage in July 2010

Motivation

- Heterogeneity of failures
  - Heterogeneity of components
    - CPU, disk, memory, network
    - Multiple network types, storage, processors in same system
- Heterogeneity of applications
  - CPU versus memory versus data intensive
- Current failure models assume one-size-fits all

Goal

- Design application-centric failure model
  - Considering component failure models
  - Considering component usage
- Implication: improve efficiency and efficacy of fault-tolerance algorithms

Approach

- Analyze 5 years of event logs of production HPC system
- Develop failure model considering heterogeneous components failures and their correlation
- Apply model to improve effectiveness of scheduling of checkpoints

Studied System

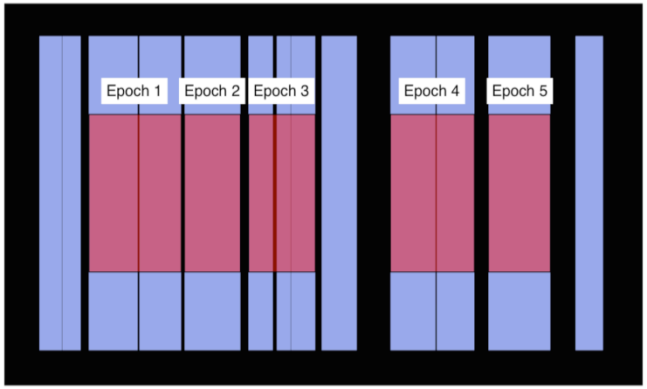
- Mercury cluster at the National Center for Supercomputing Applications (NCSA)
- Used for TeraGrid
- System evolved over time
- Cluster specification:

Resource	Phase I	Phase II
# of Nodes	256	635
Processors	2x Itanium II @ 1.3 GHz	2x Itanium II @ 1.5 GHz
Memory	4 or 12 GB DDR1600 ECC RAM	4GB DDR2100 ECC RAM
Network Storage	AFS, NFS (1TB), GPFS (90TB)	
Local Storage	1x18GB, 1x73 GB UltraSCSI drives	2x73 GB UltraSCSI drives
Network	Gigabit Ethernet, Myrinet, Management Network (Ethernet)	

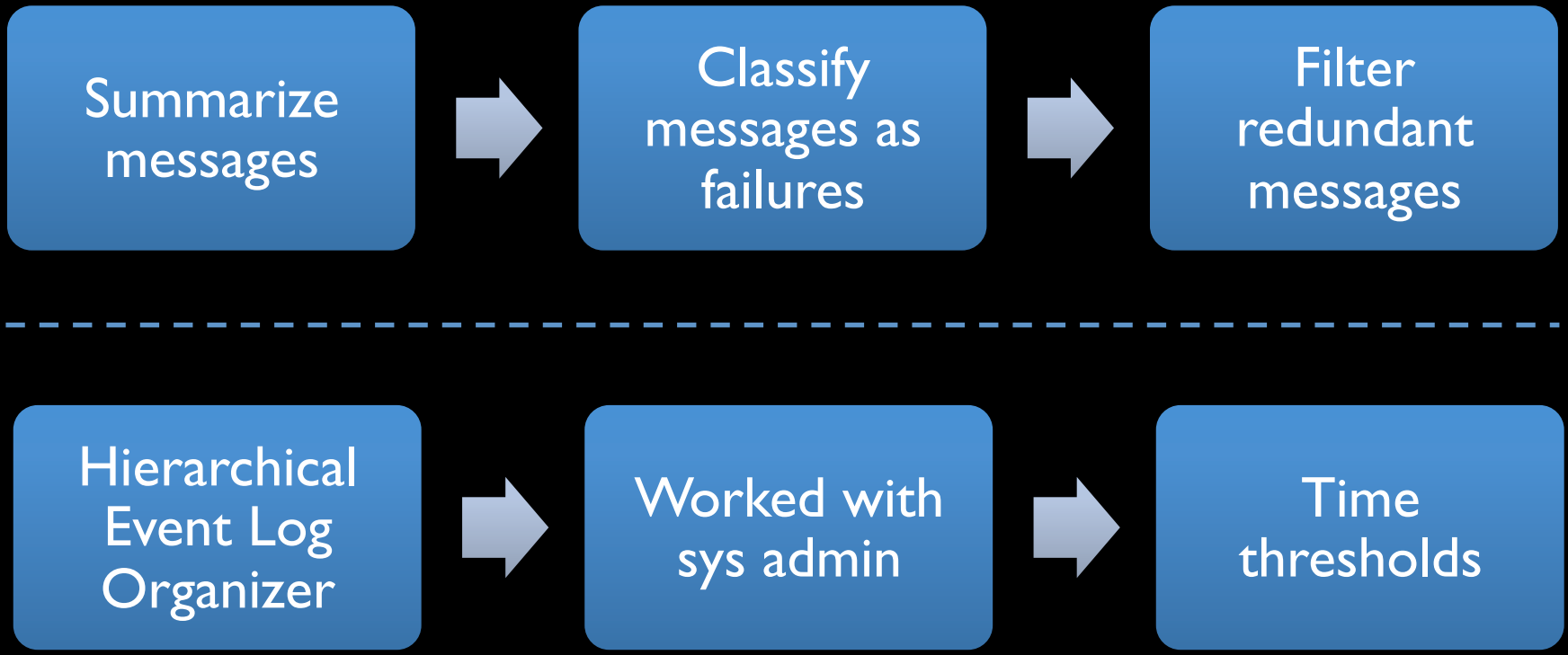
Log Collection

- Nodes' logs collected centrally
  - time of message, node, app info
- Measurement time frames over 5 years

Epoch	Time Span	Days Recorded	Days Missing
1	Jul. 2004 to Jun. 2005	329	6
2	Jun. 2005 to Jan. 2006	203	0
3	Feb. 2006 to Oct. 2006	227	15
4	Jul. 2007 to May 2008	301	4
5	Jun. 2008 to Feb. 2009	225	0
Total		1285	25



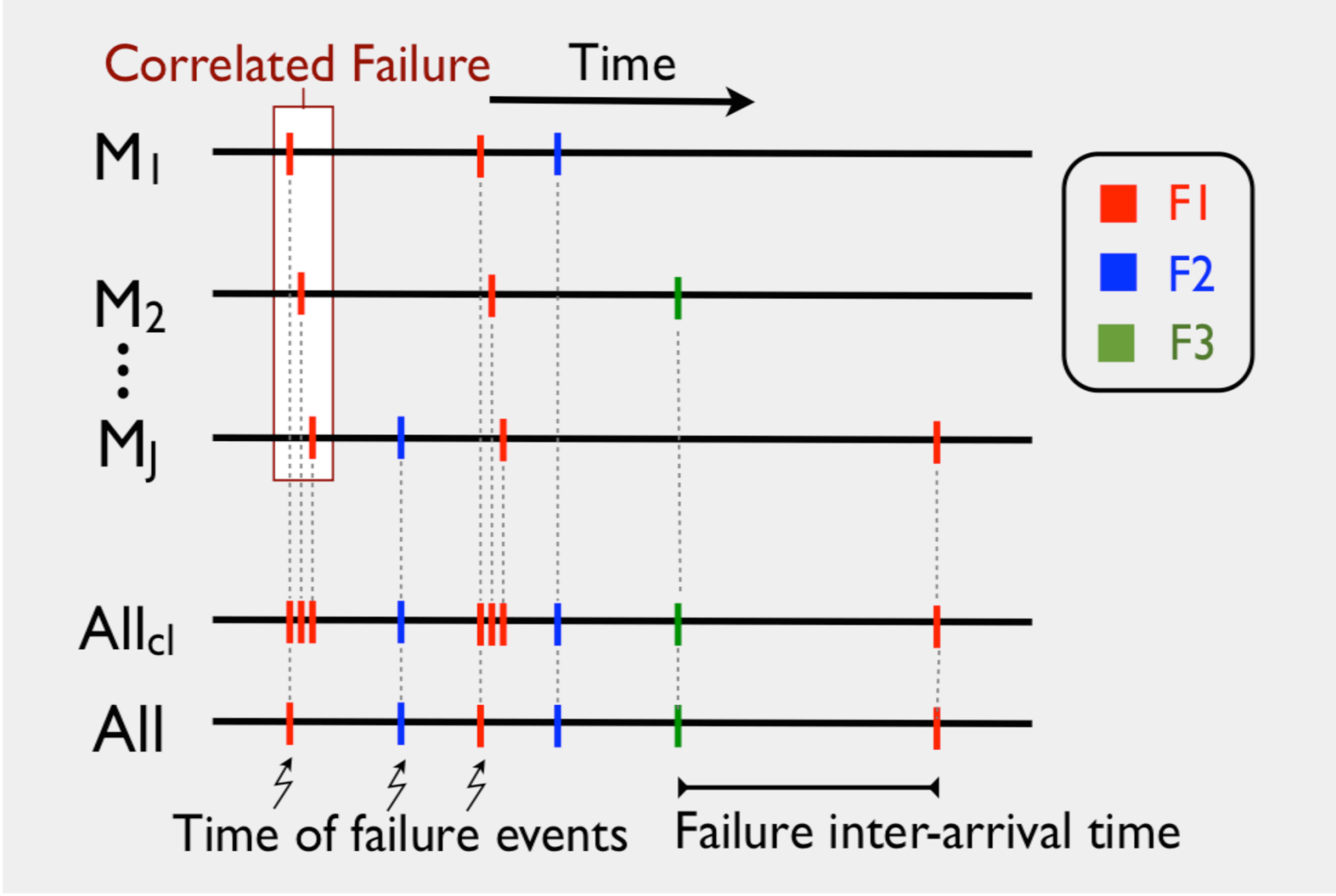
Event Processing



Error Messages

Code	Message
F1	scsi error: (1:0:0) status=02h key=4h (hardware error); fru=02h asc/ascq=11h/00h ""
F2	rpc: bad tcp reclen 0x47455420 (non-terminal)
F3	pbs_mom: sister could not communicate (15059) in xxxxxx, job_start_error from node xxxxx in job_start_error
F4	ifup: could not get a valid interface name: -> skipped
F5	+ mem error detail: physical address: 0x5fa56180, address mask: 0xfffffffff80, node: 0, card: 0, module: 4, bank: 2, device: 0, row: 6098, column: 3252
F6	processor error map: 0x4000 processor state param: xxx processor lid: 0xc0180000

Failure Events



Failure Analysis and Modeling

- Heterogeneous failure statistics
- Per-component failure model
- Integrated failure model

Heterogeneous Failure Statistics

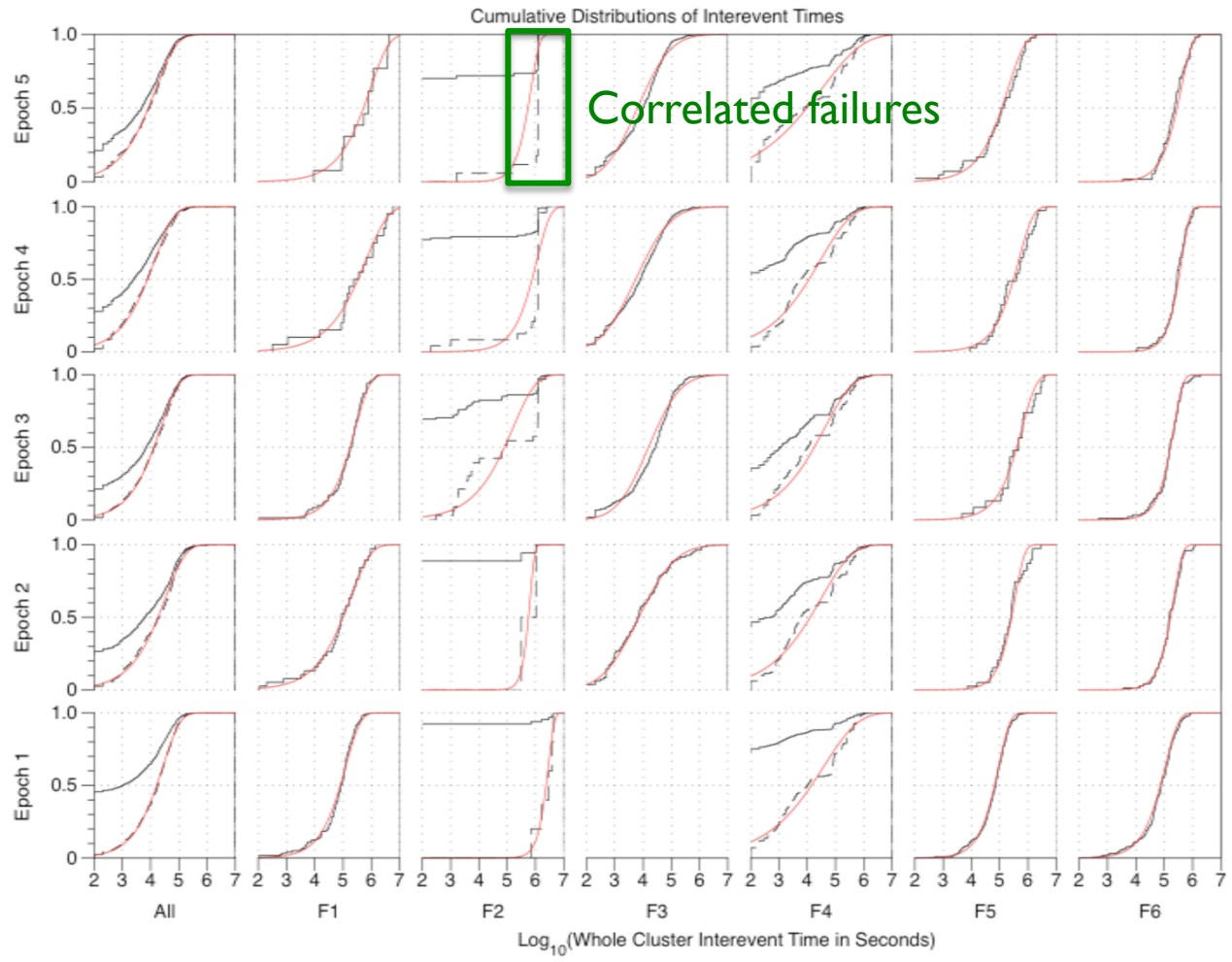
- Mean and median time to failure
- Cumulative distribution of failure inter-arrivals
- Correlation across nodes

Failure Inter-event Statistics (days)

		E1	E2	E3	E4	E5
All	Mean	0.390	0.543	0.373	0.278	0.287
	Median	0.210	0.236	0.157	0.089	0.103
F1	Mean	1.45	2.72	3.27	12.7	14.9
	Median	0.986	1.01	2.23	3.46	9.08
F2	Mean	31.2	7.90	6.90	12.2	12.4
	Median	34.6	7.90	0.778	14.0	14.0
F3	Mean	N/A	1.10	0.828	0.425	0.407
	Median	N/A	0.078	0.316	0.093	0.114
F4	Mean	1.45	1.45	1.50	1.21	1.80
	Median	0.120	0.068	0.121	0.073	0.090
F5	Mean	1.18	5.11	9.88	7.95	3.39
	Median	0.841	2.89	5.42	4.41	1.51
F6	Mean	1.52	2.65	2.68	4.27	4.09
	Median	0.908	1.74	1.82	3.25	2.70

- Average between 1.8 – 3.6 failures per day. Mean rate of between 248-484 days to failure.
- Wide range in inter-event times.

Cumulative Distribution of Inter-event Times



Whole Cluster

Discount simul-taneous failures

Best fit

Correlation

- Across nodes
  - 30-40% of the F4 failures on different machines occur within 30 seconds of each other
  - F1, F5, F6 show no correlation
- Across components
  - Divided each epoch into hour-long periods
  - Value of period is 1 if failure occurred; 0 otherwise
  - Cross-correlation: 0.04 – 0.11 on average
  - No auto-correlation either

Per-Component Failure Model

- Time to failure
  - Use Maximum Likelihood Estimation (MLE) to fit candidate distributions
  - Consider: Weibull, log-normal, log-gamma and exponential
  - Use p-values to evaluate whether empirical data could not have come from the fitted distribution

Parameters of Time to Failure Model

λ : scale, k: shape

Interevent Time (days)						
	Distribution	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
All	Weibull	$\lambda = 0.3166$ $k = 0.7097$	$\lambda = 0.3868$ $k = 0.6023$	$\lambda = 0.2613$ $k = 0.6629$	$\lambda = 0.1624$ $k = 0.6161$	$\lambda = 0.1792$ $k = 0.5841$
F1	Weibull	$\lambda = 1.436$ $k = 0.8300$	$\lambda = 2.110$ $k = 0.6052$	$\lambda = 3.167$ $k = 0.8418$	$\lambda = 7.579$ $k = 0.5560$	$\lambda = 10.684$ $k = 0.6510$
F2	Weibull	$\lambda = 33.16$ $k = 1.994$	$\lambda = 7.515$ $k = 2.631$	$\lambda = 1.831$ $k = 0.5317$	$\lambda = 13.08$ $k = 0.9249$	$\lambda = 8.077$ $k = 1.416$
F3	Log Normal	N/A	$\mu = -2.509$ $\sigma = 2.361$	$\mu = -1.717$ $\sigma = 2.030$	$\mu = -2.767$ $\sigma = 2.249$	$\mu = -2.622$ $\sigma = 2.125$
F4	Weibull	$\lambda = 0.4498$ $k = 0.3593$	$\lambda = 0.3639$ $k = 0.4009$	$\lambda = 0.4776$ $k = 0.4317$	$\lambda = 0.3400$ $k = 0.3931$	$\lambda = 0.3792$ $k = 0.2979$
F5	Weibull	$\lambda = 1.071$ $k = 1.065$	$\lambda = 4.032$ $k = 1.253$	$\lambda = 7.181$ $k = 0.8464$	$\lambda = 5.506$ $k = 0.8510$	$\lambda = 2.274$ $k = 0.7092$
F6	Weibull	$\lambda = 1.260$ $k = 0.9258$	$\lambda = 2.520$ $k = 1.392$	$\lambda = 2.520$ $k = 1.323$	$\lambda = 4.788$ $k = 1.455$	$\lambda = 4.548$ $k = 1.091$

- Heterogeneous scale and shape parameters
- F1, F4: hazard rate is decreasing for disk and network failures
- F5, F6: hazard rate is relatively constant or slightly increasing for memory or processor failures
- Overall: hazard rate is decreasing

Per-Component Failure Model

- Number of nodes in failure
  - F2, F4 can occur simultaneously on multiple nodes
    - E.g. 91% failures affect just one node, 3% affect two nodes, and 6% affect more than two nodes
- Fit distributions to number of nodes affected for the combined failures, and for F2 and F5 individually

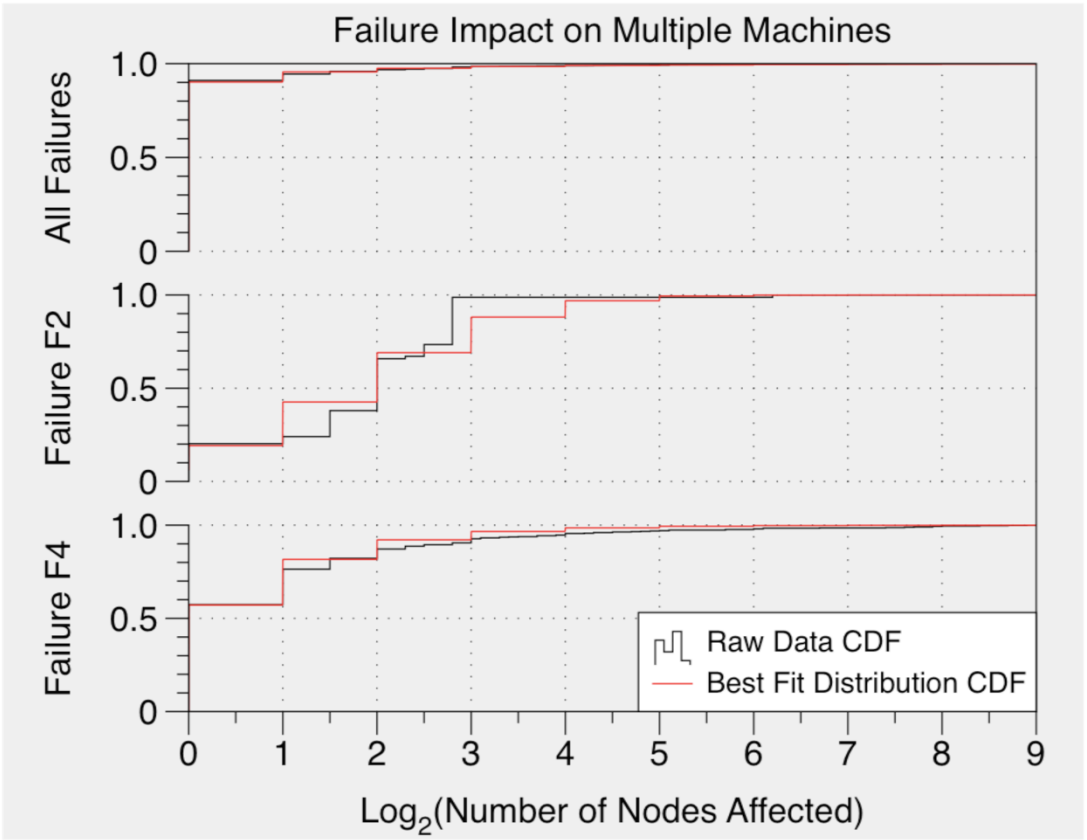
Parameters for Number of Nodes Failed Model

λ : scale, k: shape, μ : mean, σ : std dev in log-space

	Distribution	Parameters
All	Weibull	$\lambda = 0.1387, k = 0.4264$
F1, F3, F5, F6	Constant	1
F2	Log Normal	$\mu = 2.273, \sigma = 2.137$
F4	Exponential	$\lambda = 0.8469$

- Different distributions (some memoryless, others memoryfull) can fit distribution of number of nodes involved in failure

Number of Nodes in Failure



- Combined failure model
  - 91% of failure affect one node
  - 3% affect two nodes
  - 6% affect more than 2 nodes

Model Integration

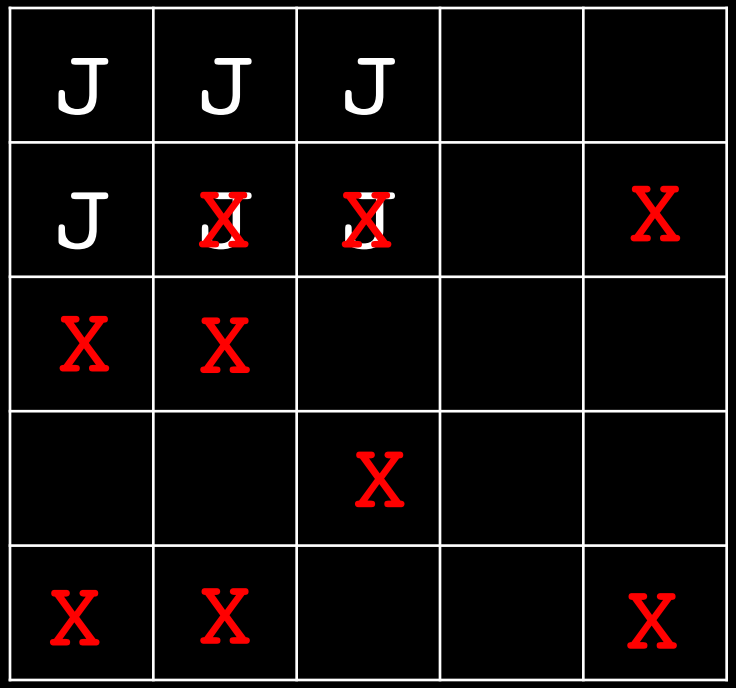
- Applications types
  - Bag-of-tasks
    - Mainly uses CPU, little memory, disk, network
    - Example: Rosetta
  - Data-intensive
    - Mainly uses disk, memory, and processor
    - Example: CM1
  - Combined
    - Corresponds to massively parallel applications
    - Uses all resources
    - Example: NAMD

Model Integration

- Holistic failure model
  - Types of components used
  - Number of nodes used
- Assumptions
  - Checkpoint occurs simultaneously across all nodes
  - Checkpointing does not affect failure rates
  - Failure on one node used by job means all job nodes must restart from checkpoint

Probability of Node Failure F_i

System with 5x5 nodes



J: Node occupied by job J
X: Node with component failure

- Calculate probability of a failure on a node used by job J
  - (Probability of 1-node failures) X (Probability it will affect job J)
  - + (Probability of 2-node failures) X (Probability it will affect job J)
  - …
  - + (Probability of N-node failures) X (Probability it will affect job J)

Probability of Node Failure F_i

Variable	Definition
M_J	Number of nodes used by job J
M_{tot}	Total number of nodes in cluster
$Q_i(N)$	Probability of number of nodes N affected by failure for component i
$P_{node}(M_J)$	Probability of a failure of component i on a node used by job J
$P_J(M_J, t)$	Probability of job J failing at time t

- Weighted sum of the probabilities that the failure affects one of more of the nodes used by the job
  - Assume each node has equal probability of failure (no space correlation)

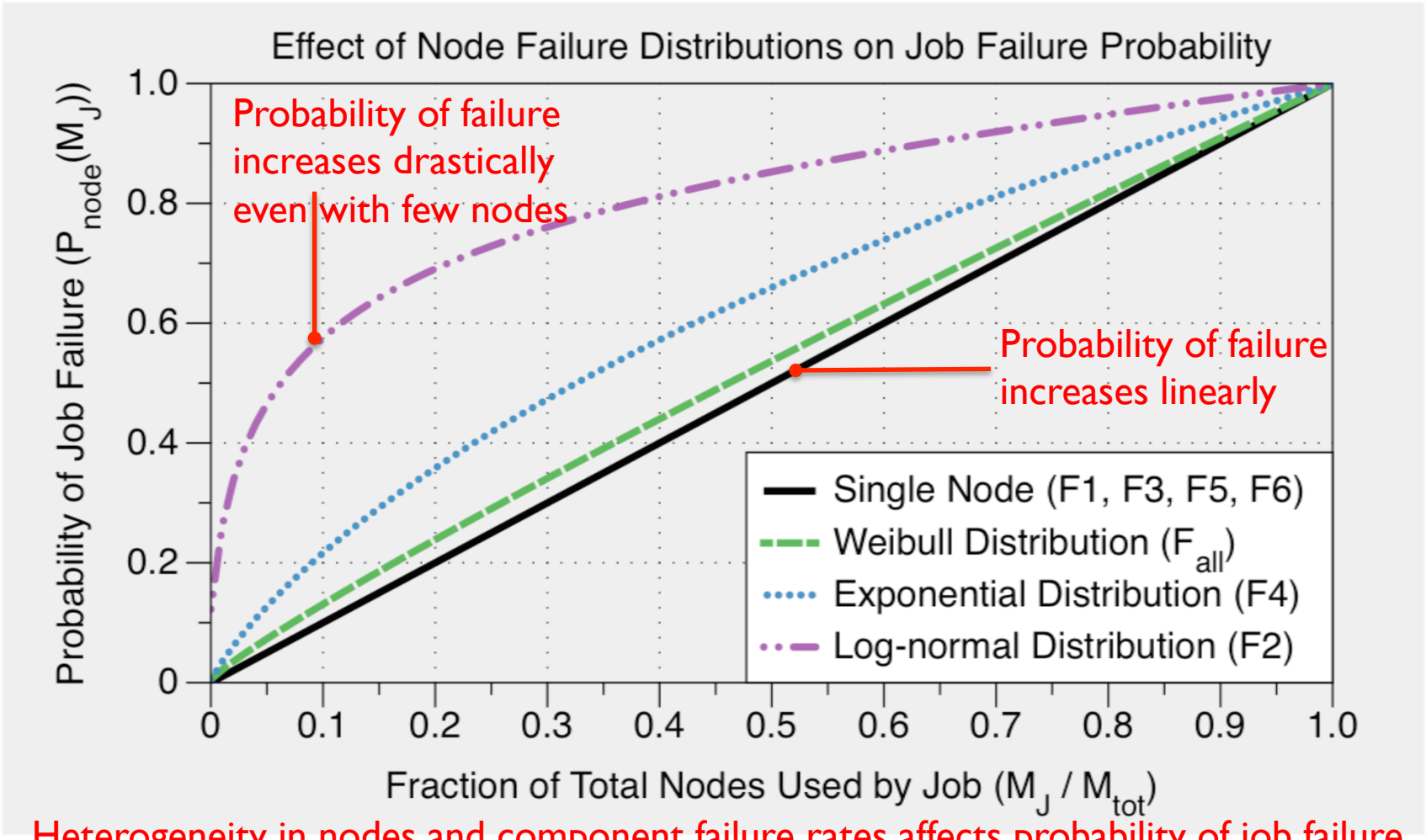
$$P_{node}(M_J) = \sum_{N=1}^{M_{tot}} Q_i(N) \left[1 - \prod_{R=0}^{N-1} \left(1 - \frac{M_J}{M_{tot} - R}\right)\right]$$

Number of nodes N that fail

Probability of N nodes failing

Probability that at least one failure affected particular node used by the job

Job Failure Probability



Heterogeneity in nodes and component failure rates affects probability of job failure

Probability of Job Failure

Variable	Definition
M_J	Number of nodes used by job J
M_{tot}	Total number of nodes in cluster
$Q_i(N)$	Probability of number of nodes N affected by failure for component i
$P_{node}(M_J)$	Probability of a failure of component i on a node used by job J
$P_J(M_J, t)$	Probability of job J failing at time t

$$P_J(M_J, t) = 1 - \prod_{i \in F_J} (1 - P_{node}(M_J) P_i(t))$$

Probability that node did not fail

Probability none of the nodes used by the job failed

Probability that at least one failure affected any one of the nodes used by the job

Compute Time to Job Failure via Monte Carlo Simulation

- For 1:num_experiments
  - time_to_fail = 0
  - while(true)
    - **time_to_fail+= Compute time to next failure**
      - Use fitted distribution for failure interarrivals
    - **prob_f = Compute probability failure will affect job**
      - Compute probability of node failing used by job J
        - » $Q_i(N)$: probability of N nodes failing
          - Used fitted distribution for number of node failures
    - **If (rand(0,1) < prob_f, store time_to_fail and exit while**

Simulation Parameters

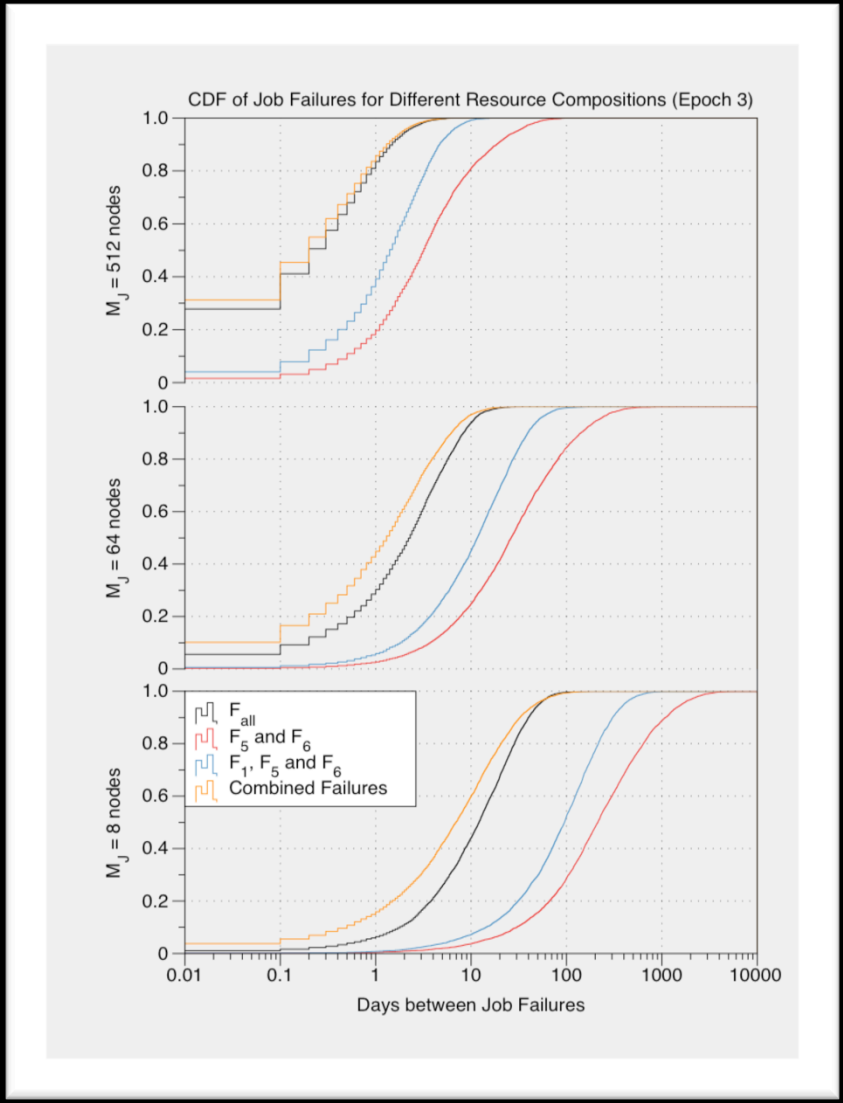
- Combinations of Failure types:
  - Combined failures
  - F_5 and F_6
  - F_1, F_5, and F_6
  - F_{all}

Failure inter-arrival times derived from fitted distributions

Failure inter-arrival times derived from traces

- Job sizes: 8, 64, 512 nodes
- Total Mercury cluster size: 891

Failure Distribution for Different Resource Usage



- Using fewer components or nodes results in longer time to failure
  - MTBF for F5, F6
    - 8 nodes: 428 days
    - 512 nodes: 7.21 days
  - MTBF with disk failures
    - 8 nodes: 135 days
    - 512 nodes: 2.13 days
- Model provides upper bound on true failure rate
  - Model overestimates number of nodes affected by F2 and F4 failures

Integrated Failure Model Parameters

Weibull Distribution of Time Between Failures (days), $M_{tot} = 891$							
	$M_J = 8$	$M_J = 16$	$M_J = 32$	$M_J = 64$	$M_J = 128$	$M_J = 256$	$M_J = 512$
F_{all}	$\lambda = 17.75$ $k = 1.013$	$\lambda = 10.68$ $k = 0.9989$	$\lambda = 6.284$ $k = 0.9915$	$\lambda = 3.379$ $k = 0.9198$	$\lambda = 1.776$ $k = 0.8860$	$\lambda = 0.9600$ $k = 0.8190$	$\lambda = 0.4765$ $k = 0.7406$
F_5, F_6	$\lambda = 363.9$ $k = 0.7167$	$\lambda = 172.8$ $k = 0.7153$	$\lambda = 89.14$ $k = 0.7180$	$\lambda = 44.15$ $k = 0.7013$	$\lambda = 21.40$ $k = 0.7270$	$\lambda = 10.90$ $k = 0.7074$	$\lambda = 5.282$ $k = 0.7020$
F_1, F_5, F_6	$\lambda = 132.7$ $k = 0.9327$	$\lambda = 65.87$ $k = 0.9867$	$\lambda = 33.52$ $k = 0.9963$	$\lambda = 16.52$ $k = 1.006$	$\lambda = 8.349$ $k = 1.049$	$\lambda = 4.225$ $k = 0.9905$	$\lambda = 2.025$ $k = 1.031$
Combined	$\lambda = 10.86$ $k = 0.7562$	$\lambda = 5.963$ $k = 0.7654$	$\lambda = 3.653$ $k = 0.8173$	$\lambda = 2.236$ $k = 0.7959$	$\lambda = 1.429$ $k = 0.7887$	$\lambda = 0.8193$ $k = 0.7513$	$\lambda = 0.4419$ $k = 0.7222$

λ : scale, k: shape

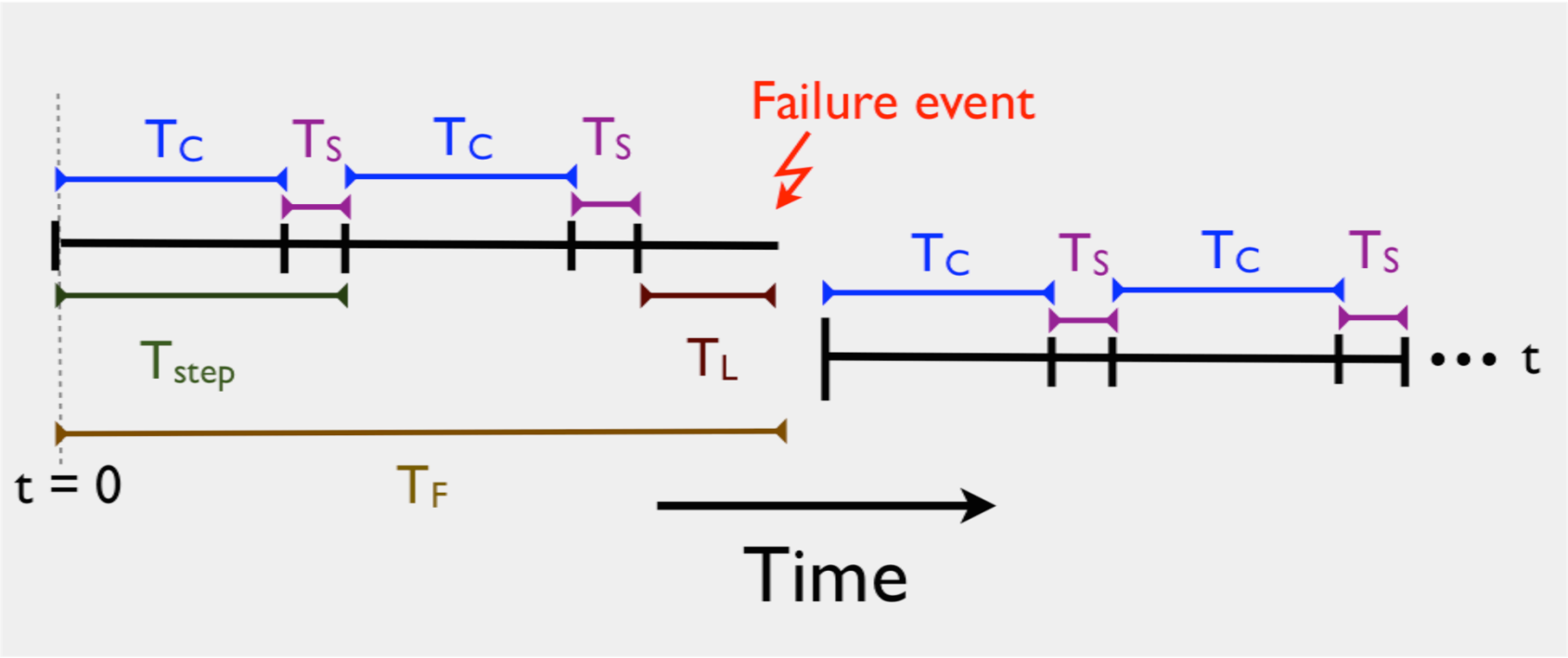
- Best fitting distribution to model is Weibull
- Scale changes as number of nodes used by job changes. Shape parameter stays mostly the same
  - Doubling number of job nodes results in halving of λ

Tolerating Heterogeneous Failures with Checkpointing

- Tradeoff between cost of checkpointing and lost work due to failures
- Novel problem
  - Checkpoint strategies that account for correlated failures and mixtures of component failure rates
- Side-effect
  - Show new models are useful

Checkpoint Model

Variable	Definition
T_C	Computation time
T_S	Time to checkpoint
T_L	Lost computation since last checkpoint
T_{step}	Single computation and checkpoint step $(T_C + T_S)$
T_F	Time from (re)start of computation to next failure
T_W	Total time wasted



Expected Wasted Time

Variable	Definition
T_C	Computation time
T_S	Time to checkpoint
T_L	Lost computation since last checkpoint
T_{step}	Single computation and checkpoint step $(T_C + T_S)$
T_F	Time from (re)start of computation to next failure
T_W	Total time wasted

- Compute T_W as a weighted average of overhead incurred for all possible times of failure
- Assume infinite computation time
- Assume constant checkpoint frequency

$$T_W = \sum_{n=0}^{\infty} \int_{nT_{step}}^{(n+1)T_{step}} [t - nT_C] \left[\frac{k}{\lambda} \left(\frac{t}{\lambda} \right)^{k-1} e^{-(\frac{t}{\lambda})^k} \right] dt$$

Segment where failure occurs

Failure between time nT_{step} to $(n+1)T_{step}$

Time wasted = total time – time computing

Weibull density function for failure occurring at time t

Optimal Checkpoint Intervals (h)

		$M_J = 8$	$M_J = 16$	$M_J = 32$	$M_J = 64$	$M_J = 128$	$M_J = 256$	$M_J = 512$
F_{all}	T_S =1 min	3.746	2.912	2.236	1.669	1.222	0.923	0.681
	T_S =10 min	11.77	9.136	6.995	5.218	3.810	2.881	2.134
	T_S =30 min	20.24	15.68	11.98	8.923	6.495	4.906	3.640
F_5, F_6	T_S =1 min	18.99	13.38	9.430	6.713	4.594	3.327	2.328
	T_S =10 min	60.17	42.42	29.90	21.30	14.57	10.56	7.392
	T_S =30 min	104.55	73.67	51.91	36.97	25.25	18.32	12.81
F_1, F_5, F_6	T_S =1 min	10.44	7.345	5.171	3.619	2.547	1.831	1.253
	T_S =10 min	33.16	23.24	16.28	11.37	7.972	5.718	3.882
	T_S =30 min	57.69	40.27	28.06	19.55	13.66	9.767	6.579
Combined	T_S =1 min	3.216	2.370	1.806	1.427	1.145	0.8866	0.6650
	T_S =10 min	10.18	7.493	5.680	4.488	3.594	2.786	2.089
	T_S =30 min	17.61	12.94	9.766	7.707	6.158	4.769	3.570

- As # of machines increases, optimal checkpoint interval increases:

$$T_C \propto 1/\sqrt{M_J}$$

- As overhead for checkpointing increases, checkpoint period increases:

$$T_C \propto \sqrt{T_S}$$

- For a given application profile, checkpoint frequency differs greatly.
  - With 64 node, checkpoint period for F_5, F_6 is ~4 times less frequent for F_{all}

Related Work

- Message log analysis
  - HELO tool shows higher accuracy
- Analysis of failures in large systems
  - Do not distinguish among component failures nor component usage
  - Or focus exclusively on a particular component type
- Fault-tolerance for parallel computing
  - Checkpointing assumes homogeneous failure model in terms of components and independence

Conclusion

- Measurement
  - Identified and analyzed five years of event logs from production HPC system
- Modeling
  - Determine distribution of failure inter-arrival times of different components
  - Formed holistic failure model, given application profile
- Fault-tolerance
  - Applied model to derive optimal time to checkpoint, considering correlated failures and heterogeneous failure rates per component

Future Work

- Measurement
  - Failure Trace Archive
- Modeling
  - Development of hierarchical failure models with space correlation
- Fault-tolerance
  - Closed-form expression for optimal checkpoint period
  - Optimal number of nodes to minimize failure rate

Probability of Node Failure F_i

Variable	Definition
M_J	Number of nodes used by job J
M_{tot}	Total number of nodes in cluster
$Q_i(N)$	Probability of number of nodes N affected by failure for component i
$P_{node}(M_J)$	Probability of a failure of component i on a node used by job J
$P_J(M_J, t)$	Probability of job J failing at time t

Weighted sum of the probabilities that the failure affects one of more of the nodes used by the job

$$P_{node}(M_J) = \sum_{N=1}^{M_{tot}} Q_i(N) \left[1 - \prod_{R=0}^{N-1} \left(1 - \frac{M_J}{M_{tot} - R}\right)\right]$$

Probability that none of the N nodes fail

Probability a node does not fail

Probability that the node fails

Number of nodes N that fail

Probability of N nodes failing

Probability that at least one failure affected particular node used by the job

Example for $M_{tot} = 4$

$$P_{node}(M_J) = Q_i(1)\,[M_j/M_{tot}] + Q_i(2)\left[1 - \left(1 - \frac{M_J}{M_{tot}}\right)\left(1 - \frac{M_J}{M_{tot} - 1}\right)\right] \cdots$$

Probability that one out of the M_j nodes will fail.

Chance that first failure will affect M_J job nodes out of M_{tot}

Chance that second failure will affect M_J job nodes out of M_{tot} – 1remaining

Chance first failure did not affect any job nodes

Chance that second failure did not affect job nodes

Chance that at least one job node affected by 2-node failures

BACKUP SLIDES