The Failure Trace Archive:
Enabling Comparative Analysis of Diverse Distributed Systems

Derrick Kondo[1], Bahman Javadi[1],
Alexandru Iosup[2], Dick Epema[2]

[1]INRIA, France [2]TU Delft, The Netherlands

Motivation

- Push toward experimental computer *science*

Motivation

- Push toward experimental computer *science*
- Hard to evaluate and compare algorithms and models for fault-tolerance
  - Lack of public trace data sets
  - Lack of standard trace format
  - Lack of parsing and analytical tools

Motivation

- Push toward experimental computer *science*
- Hard to evaluate and compare algorithms and models for fault-tolerance
  - Lack of public trace data sets
  - Lack of standard trace format
  - Lack of parsing and analytical tools
- Failures in distributed systems have increasingly high negative impact and complex dynamics

Failure Trace Archive (FTA)

http://fta.inria.fr

- Availability traces of distributed systems, differing in scale, volatility, and usage
- Standard event-based format for failure traces
- Scripts and tools for parsing and analyzing traces in svn repository

Related Work

Resource	Data Sets	Format	Parsing Tools	Analysis Tools
Grid Observatory	Emphasis on EGEE	✗	✗	✗
Computer Failure Repo.	12 (mainly clusters)	✗	✗	✗
Repo. of Avail. Traces	5 (mainly P2P)	✗✓	✓	✗
Desktop Grid Archive	4 Desktop Grids	✗✓	✗	✗
FTA[1]	22	✓	✓	✓

[1] FTA includes data sets of the former three resources, in addition to providing several new data sets

Enabled Studies

- Comparing models/algorithms using the identical data sets

Enabled Studies

- Comparing models/algorithms using the identical data sets
- Evaluation of generality/specificity of model/algorithm across different types of systems

Enabled Studies

- Comparing models/algorithms using the identical data sets
- Evaluation of generality/specificity of model/algorithm across different types of systems
- Evaluation of the generality of a system trace

Enabled Studies

- Comparing models/algorithms using the identical data sets
- Evaluation of generality/specificity of model/algorithm across different types of systems
- Evaluation of the generality of a system trace
- Analysis of evolution of failures over time

Enabled Studies

- Comparing models/algorithms using the identical data sets
- Evaluation of generality/specificity of model/algorithm across different types of systems
- Evaluation of the generality of a system trace
- Analysis of evolution of failures over time
- And many more...

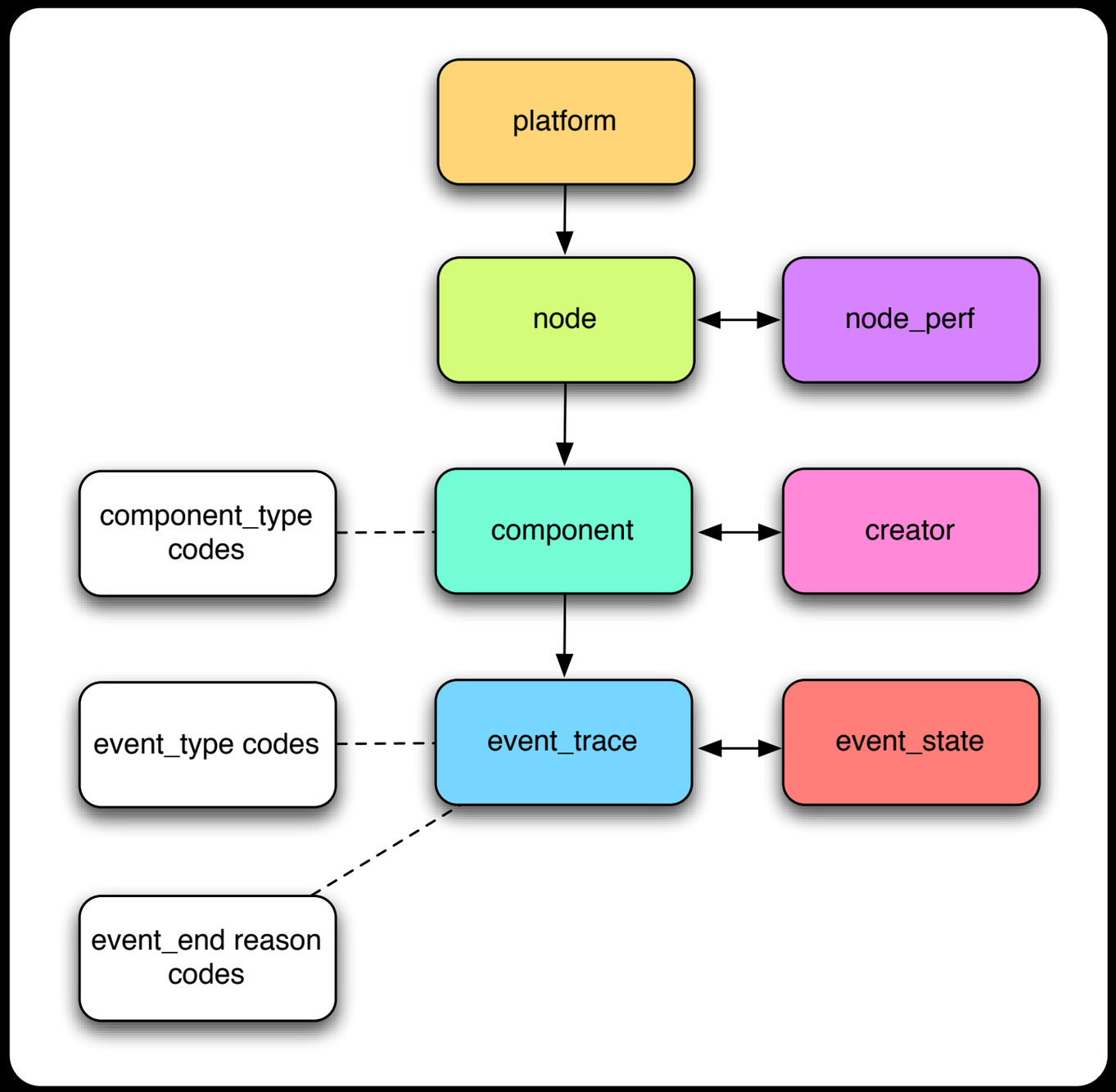
Contributions

- Description of FTA, trace format and analysis toolbox
- High-level statistical characterization of failures in each data set
- Show importance of public data sets and methods via characterization of ambiguous data sets

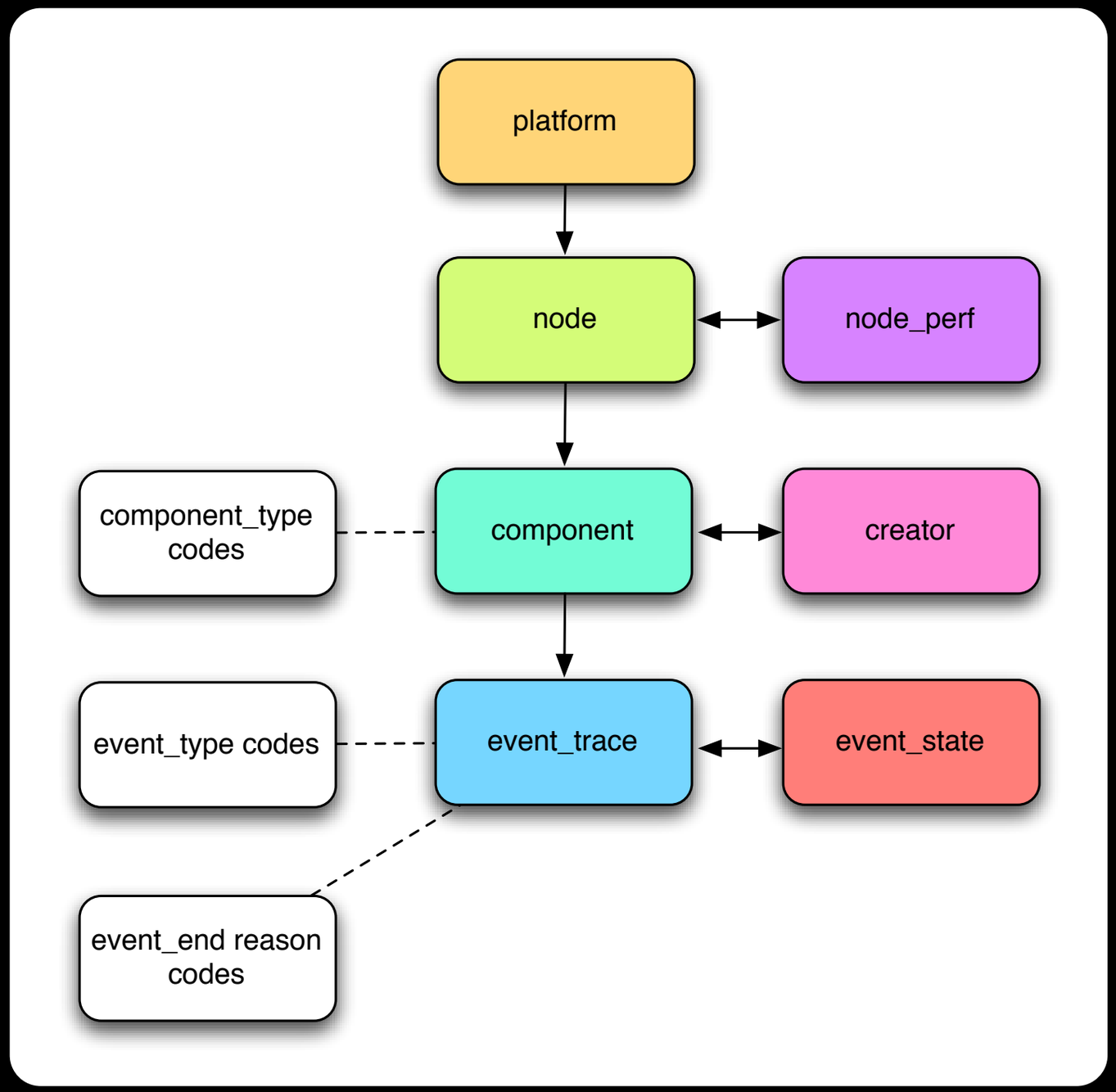
Background Definitions

- Failure: observed deviation from correct system state
  - Availability (unavailability) interval: continuous period that system is in correct state (incorrect state)
- Error: system state (not externally visible) that leads to failure
- Fault: root cause of an error

FTA Schema

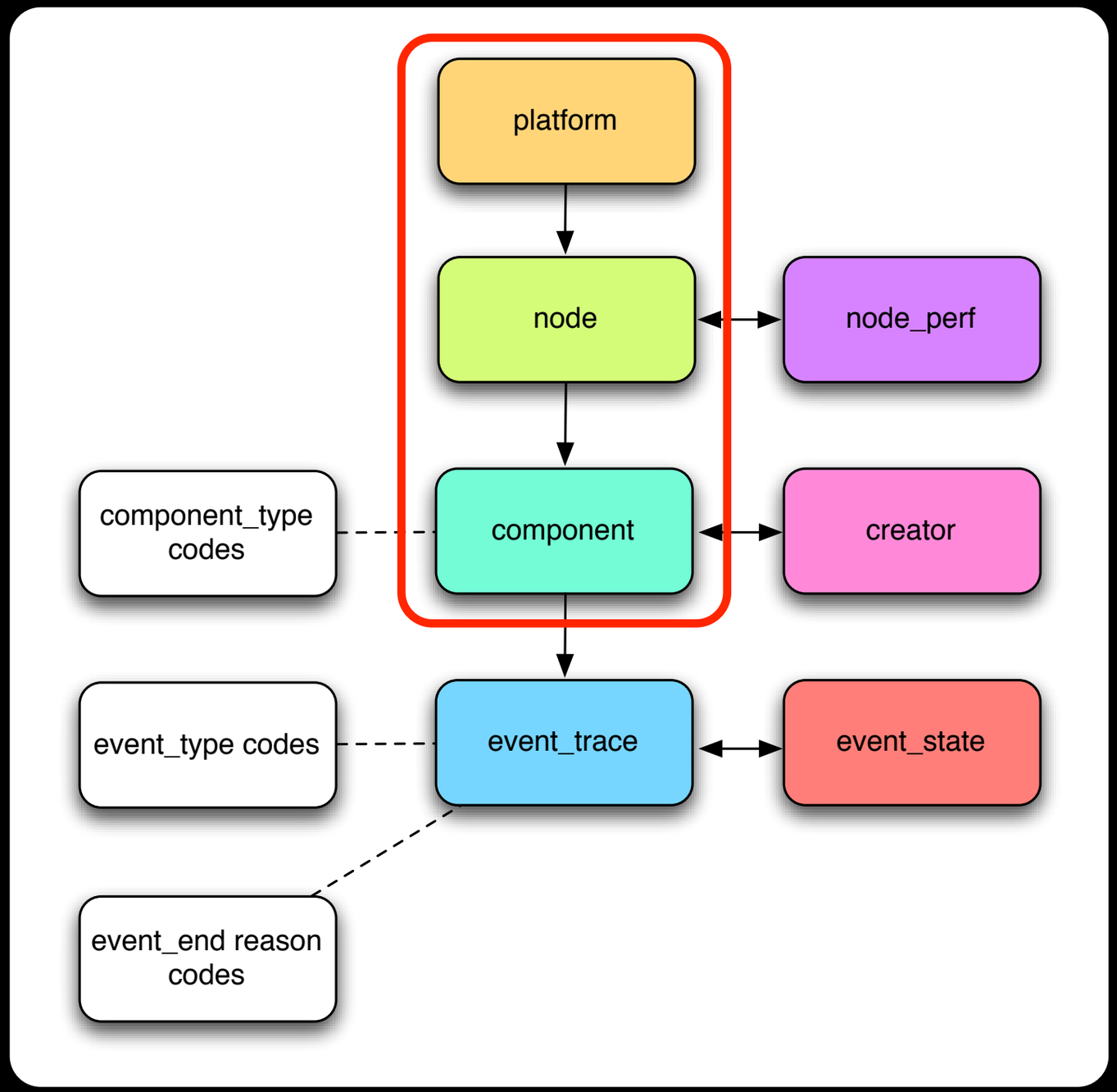

platform
node
node_perf
component_type codes
component
creator
event_type codes
event_trace
event_state
event_end reason codes

FTA Schema



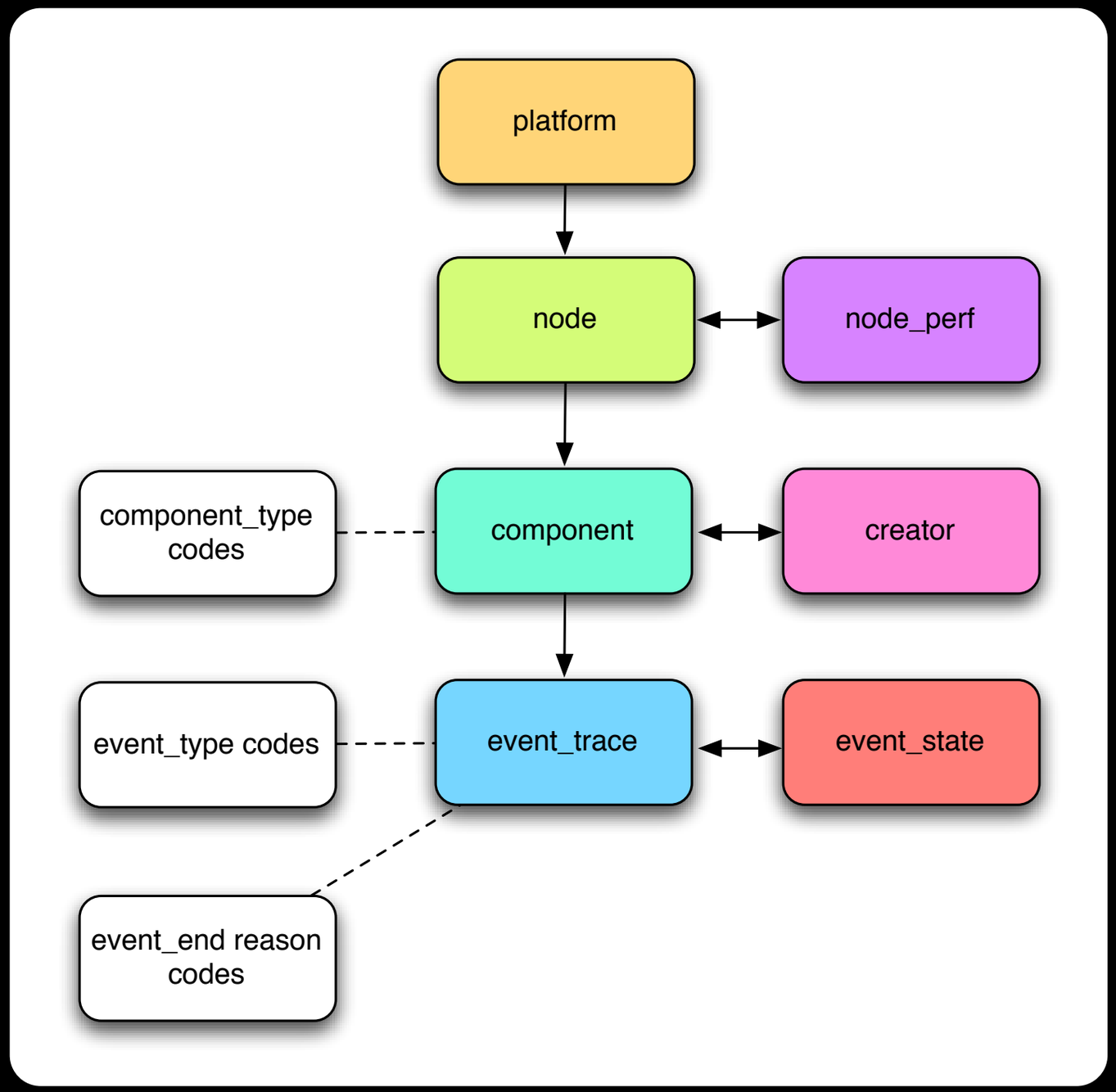
- Resource (versus job or user) centric

FTA Schema



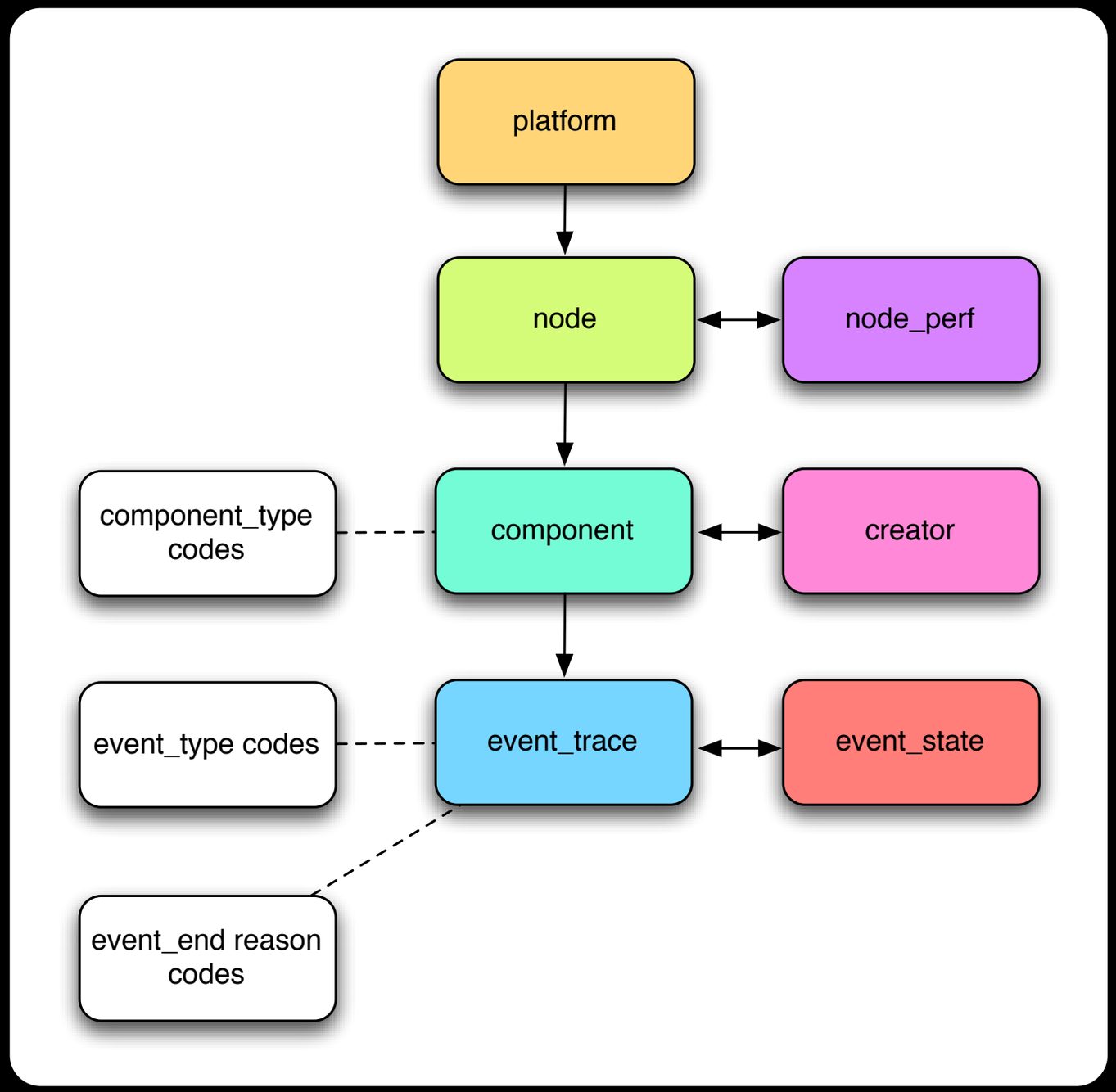
- Resource (versus job or user) centric

FTA Schema



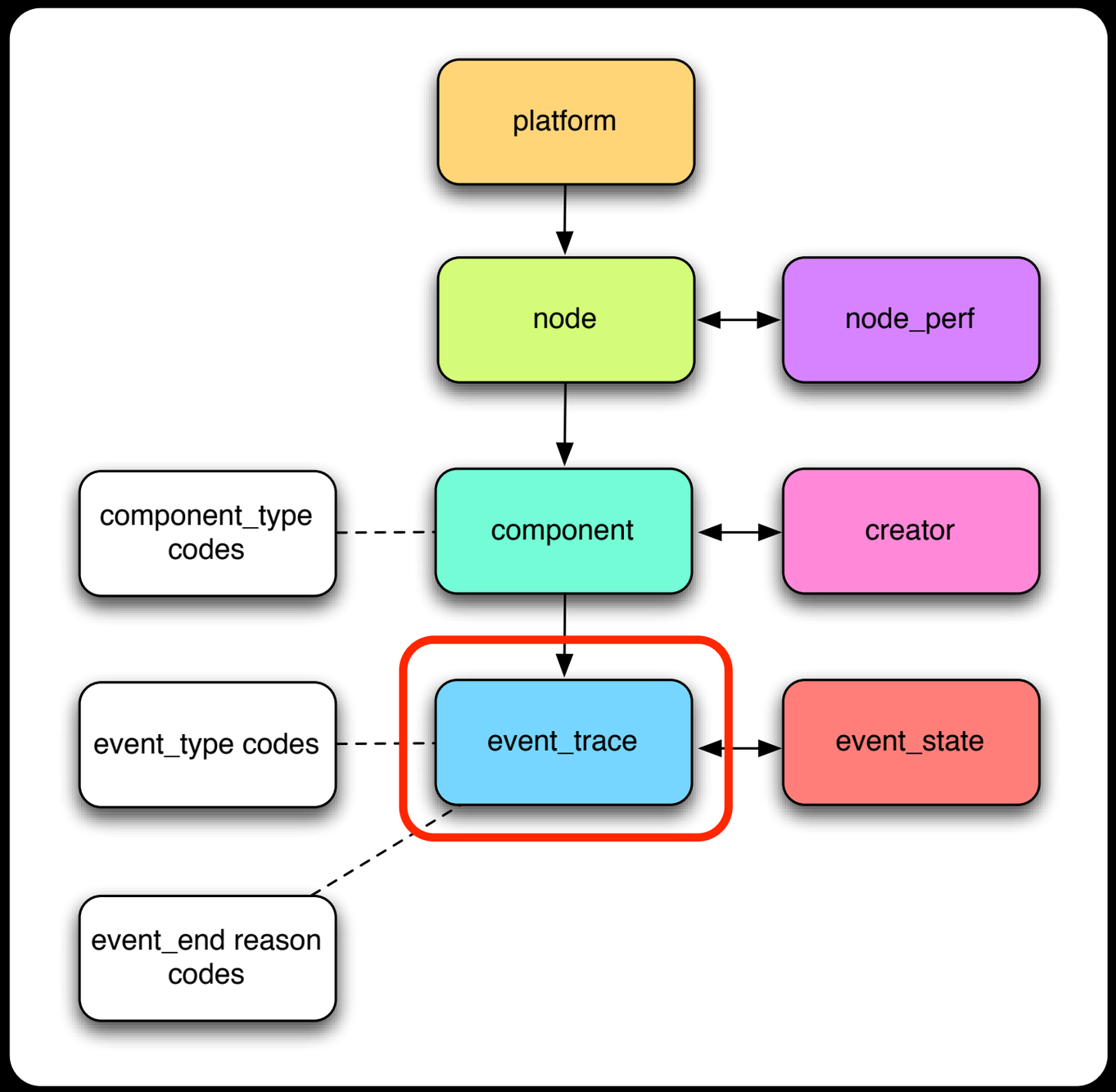
- Resource (versus job or user) centric

FTA Schema



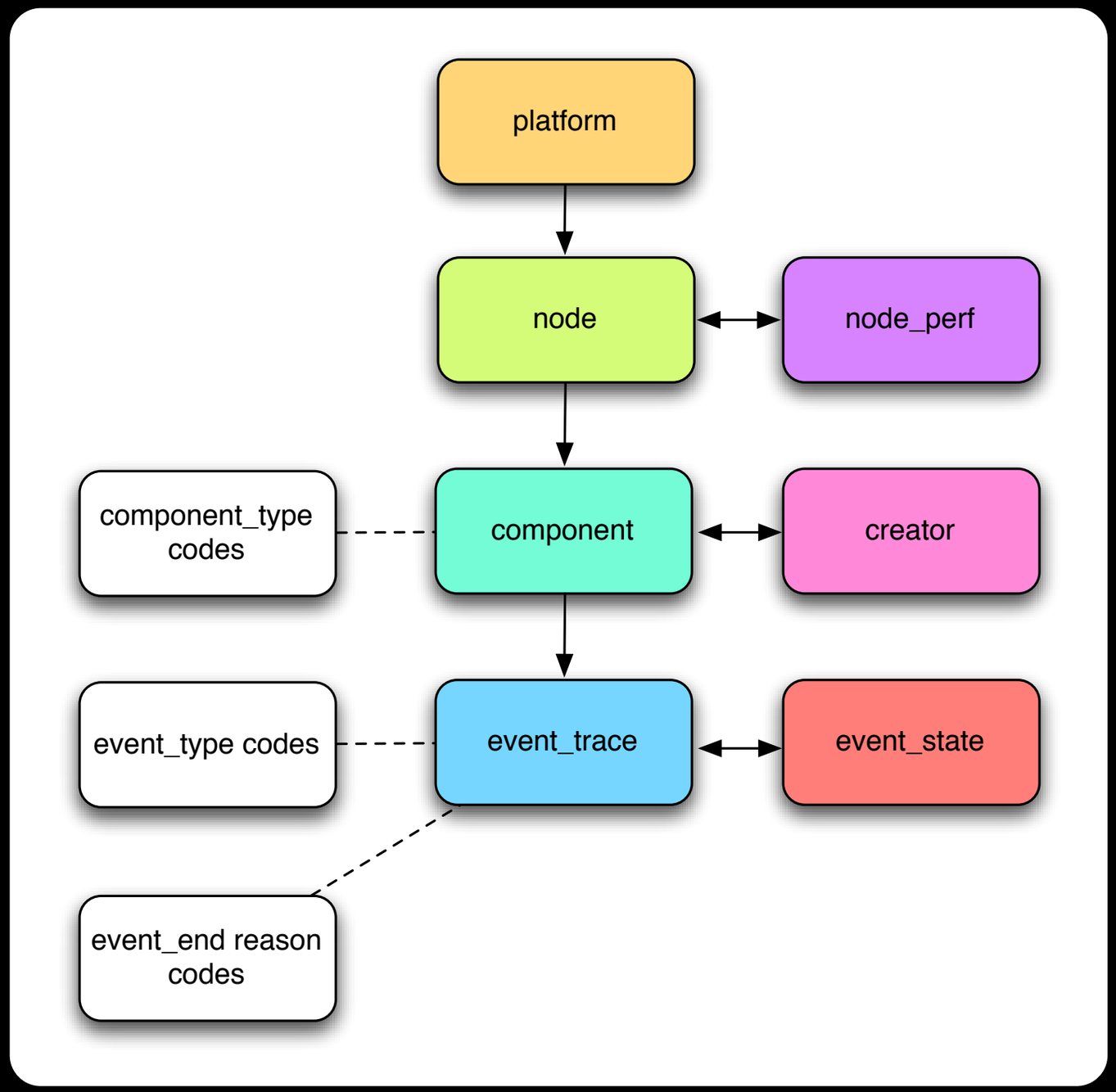
- Resource (versus job or user) centric
- Event-based

FTA Schema



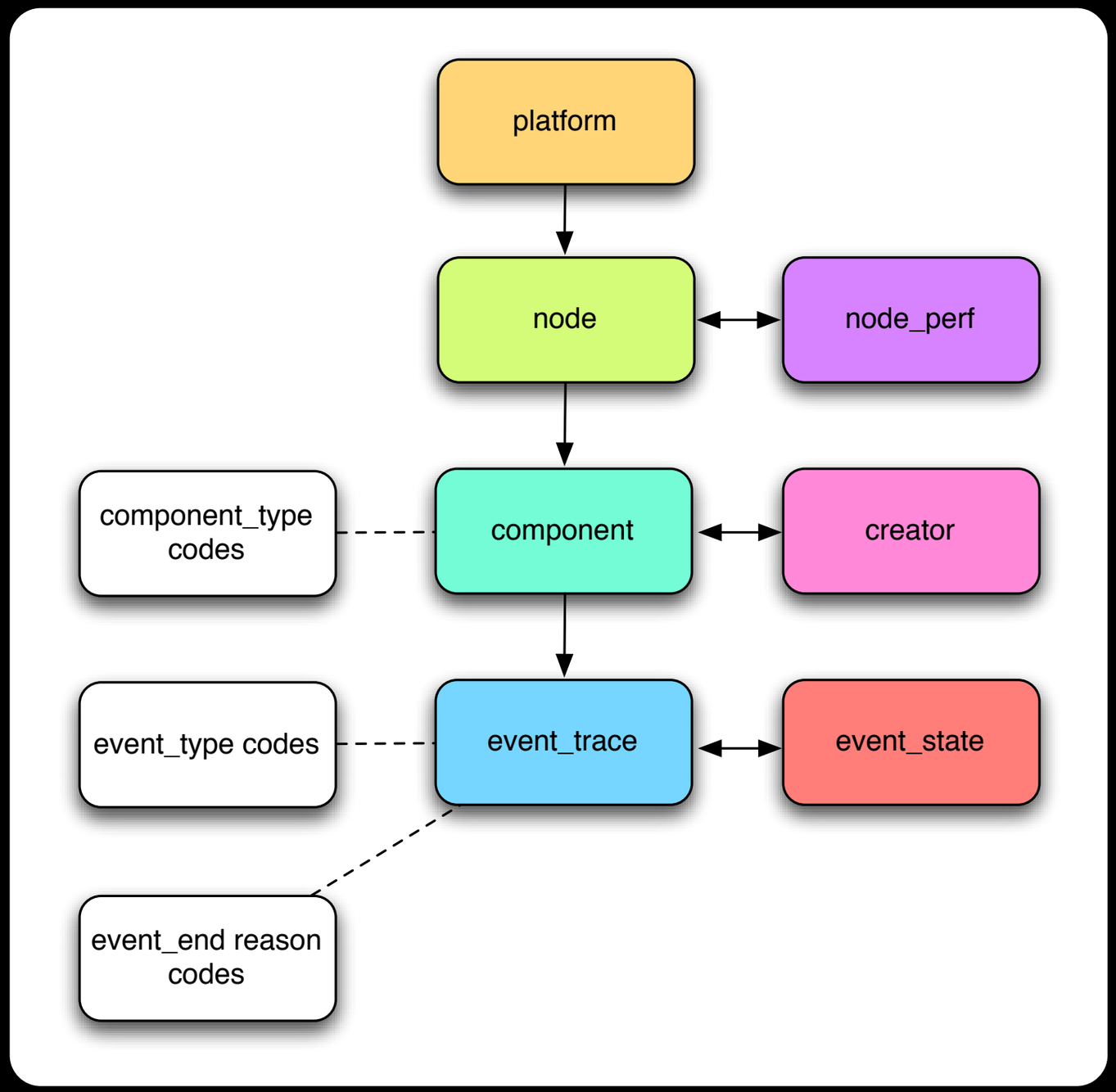
- Resource (versus job or user) centric
- Event-based

FTA Schema



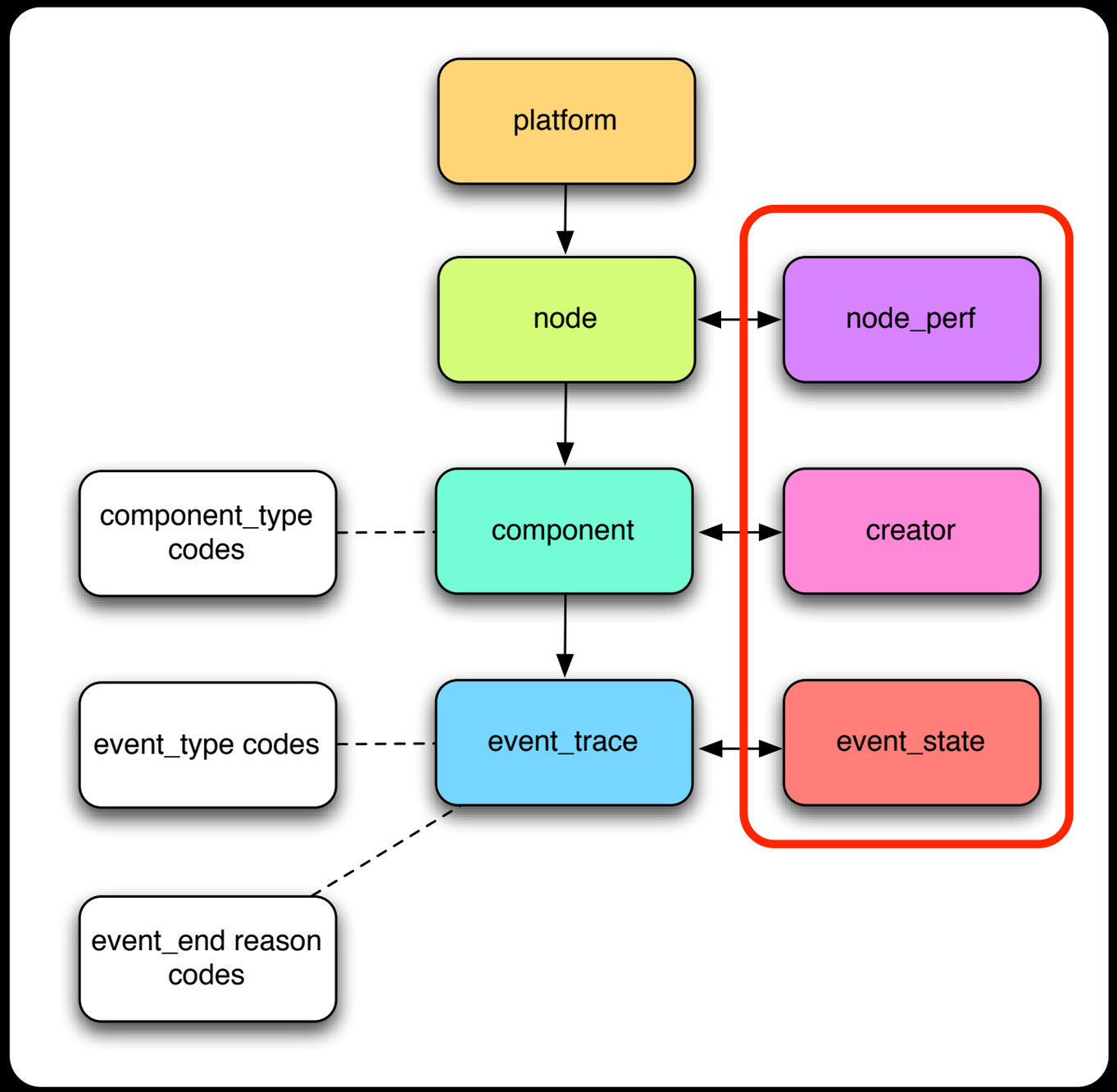
- Resource (versus job or user) centric
- Event-based

FTA Schema



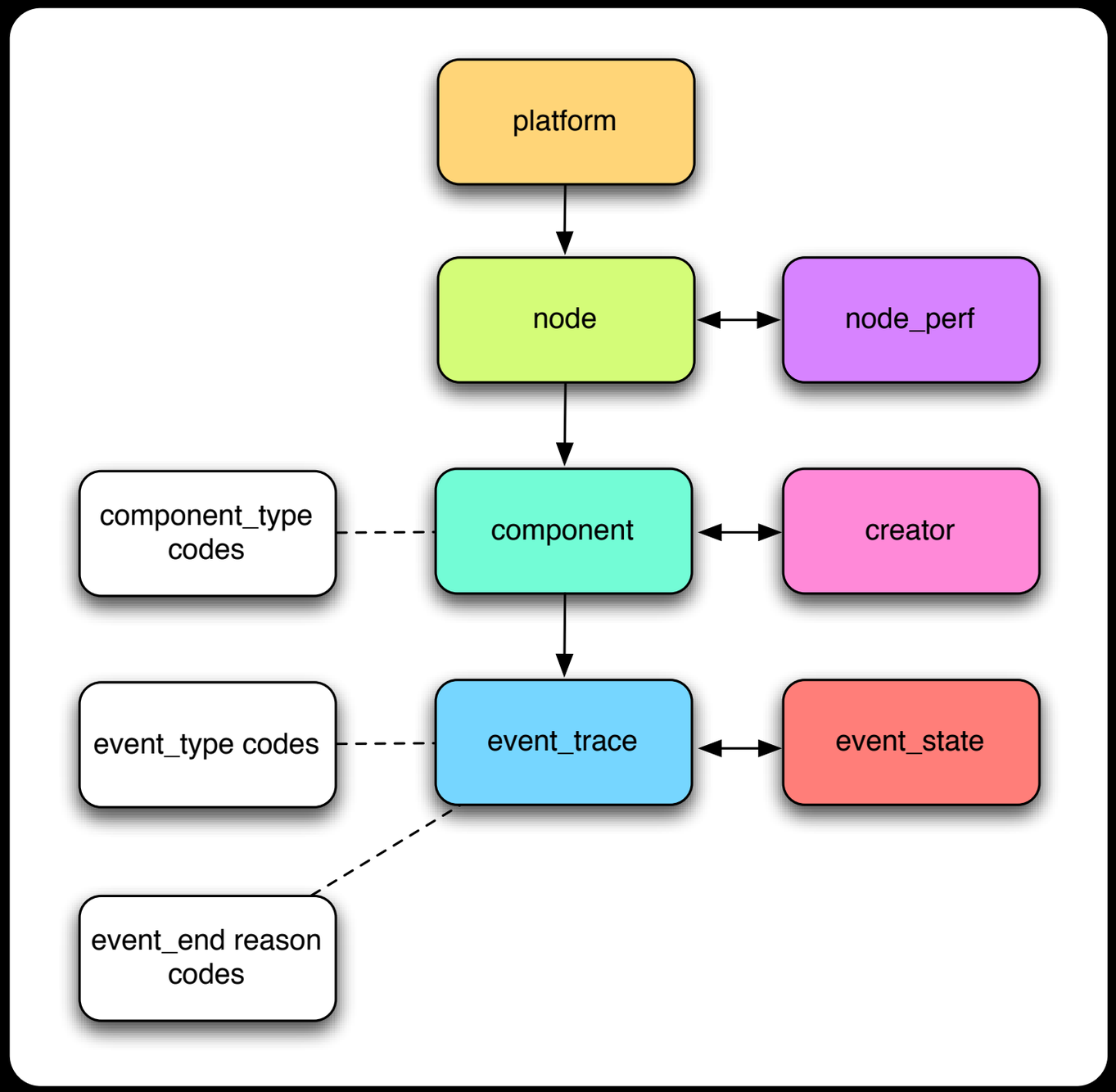
- Resource (versus job or user) centric
- Event-based
- Associated metadata

FTA Schema



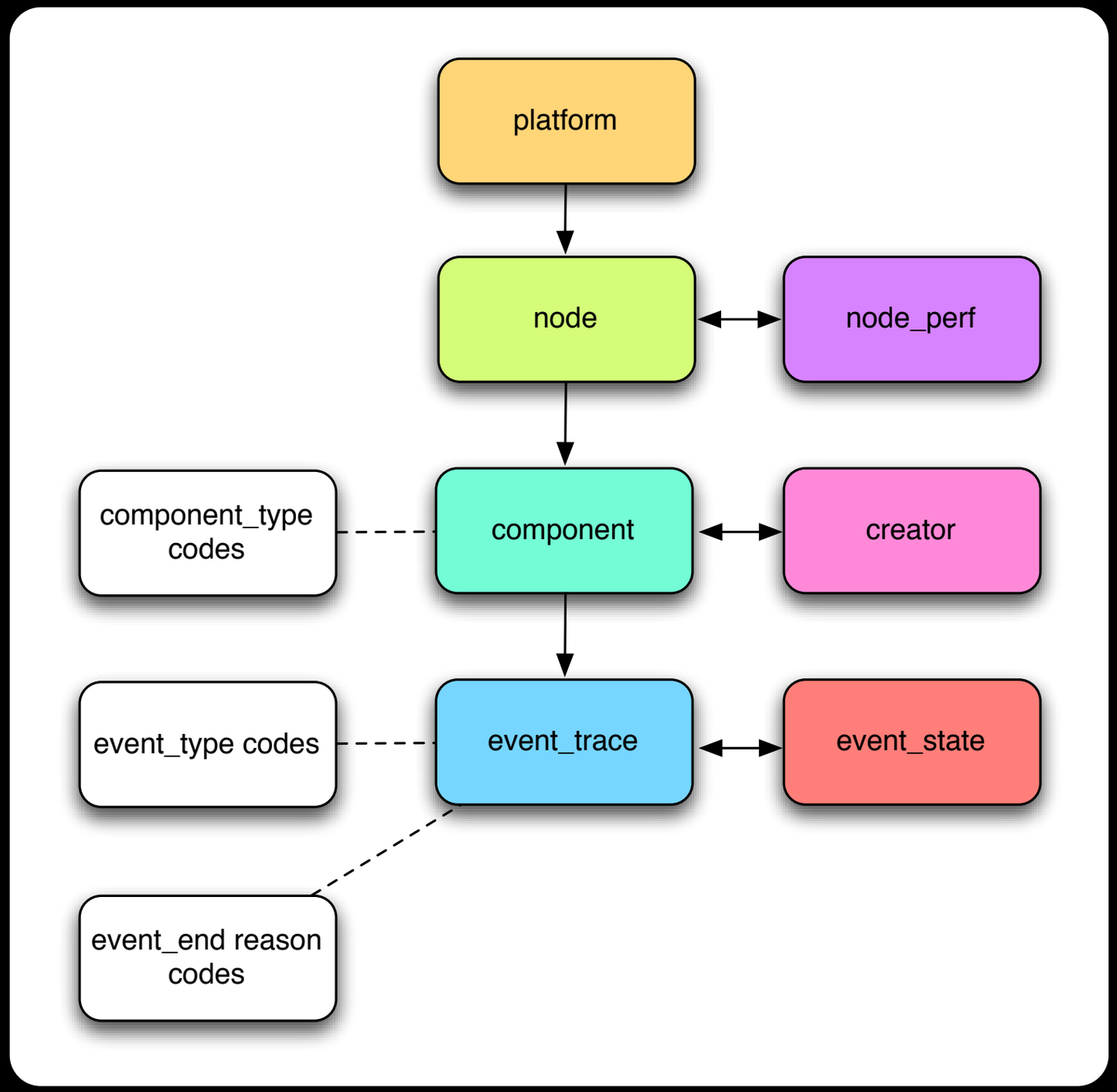
- Resource (versus job or user) centric
- Event-based
- Associated metadata

FTA Schema



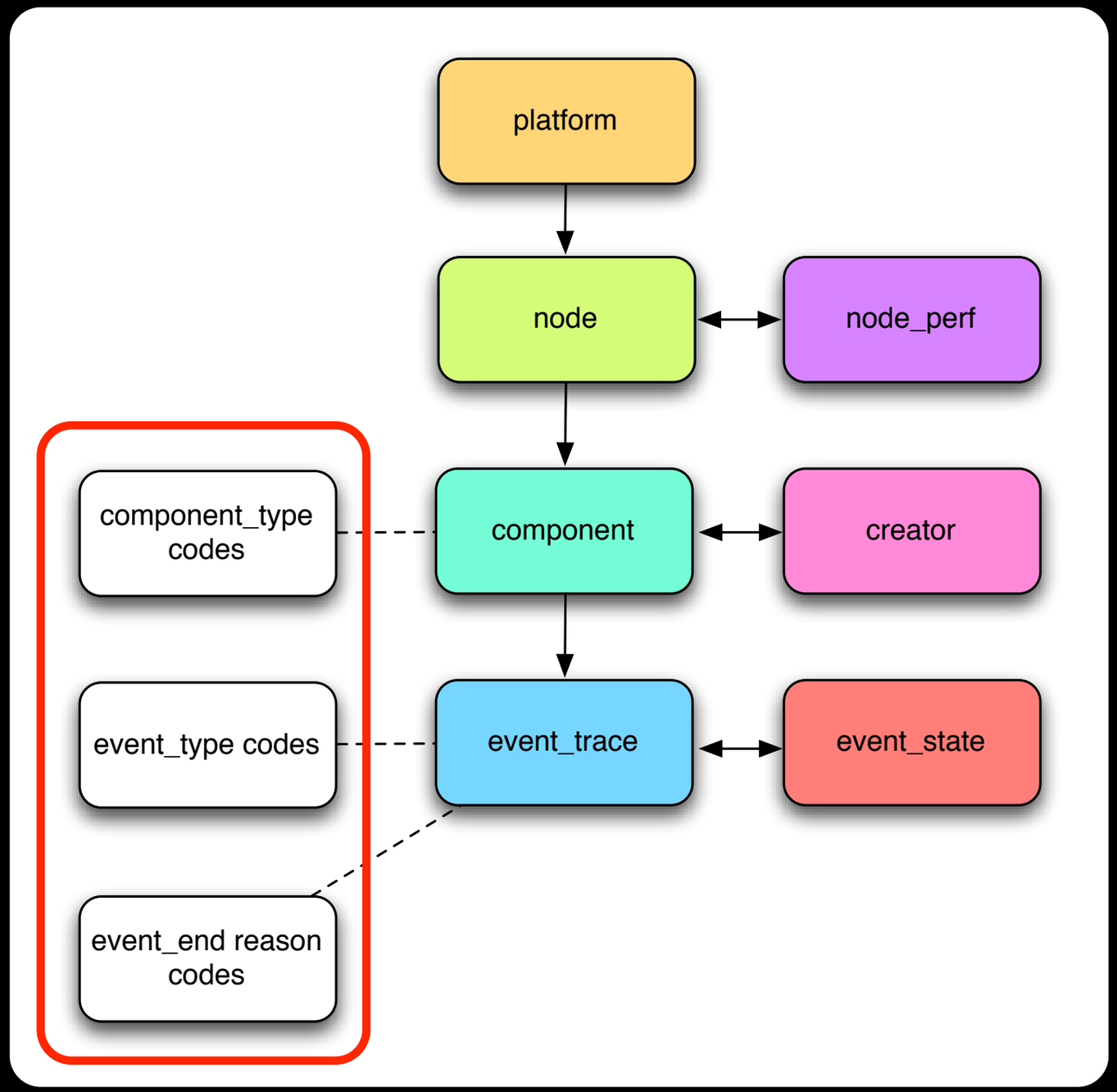
- Resource (versus job or user) centric
- Event-based
- Associated metadata

FTA Schema



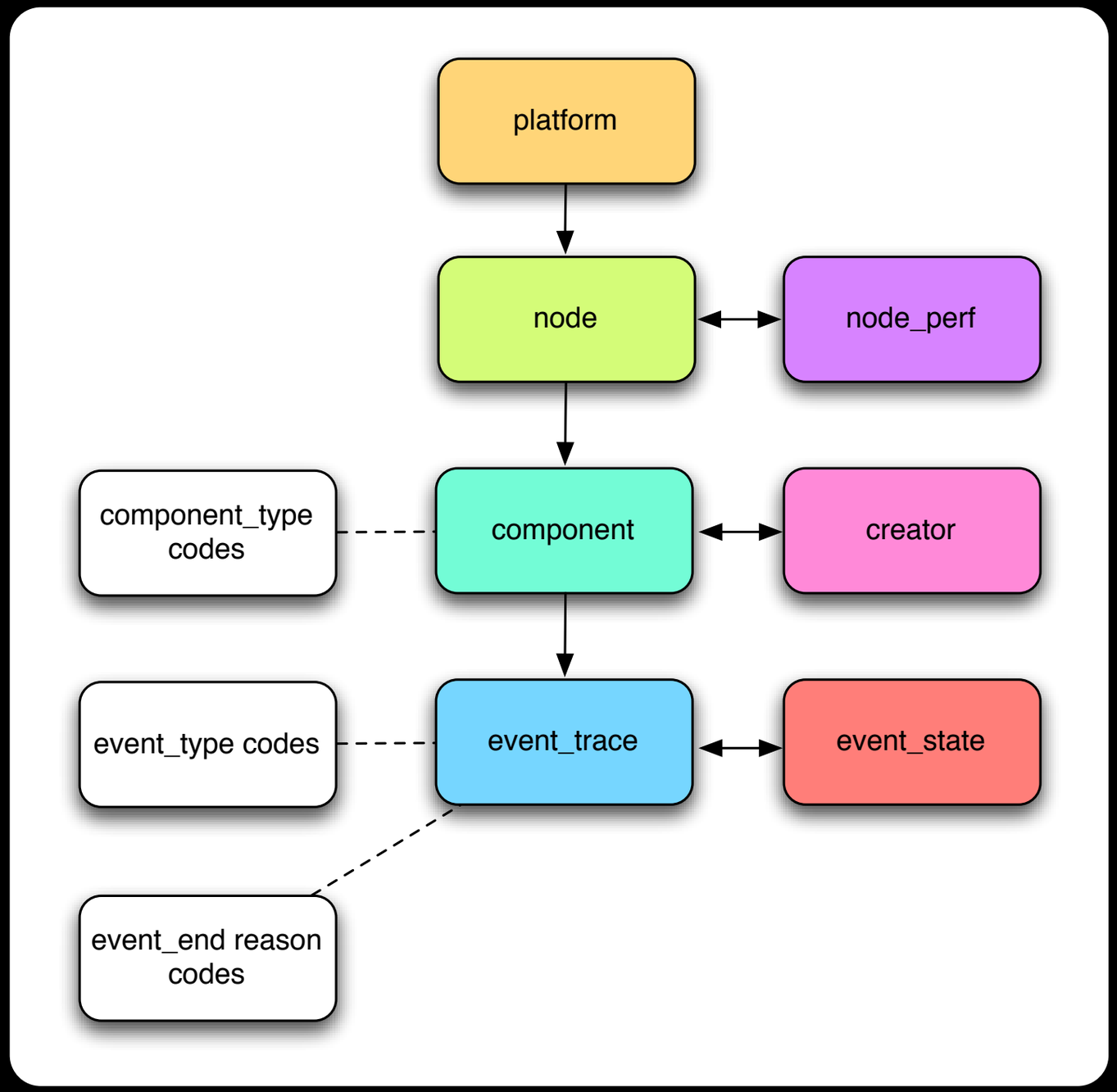
- Resource (versus job or user) centric
- Event-based
- Associated metadata
- Codes for different components, events, and errors

FTA Schema



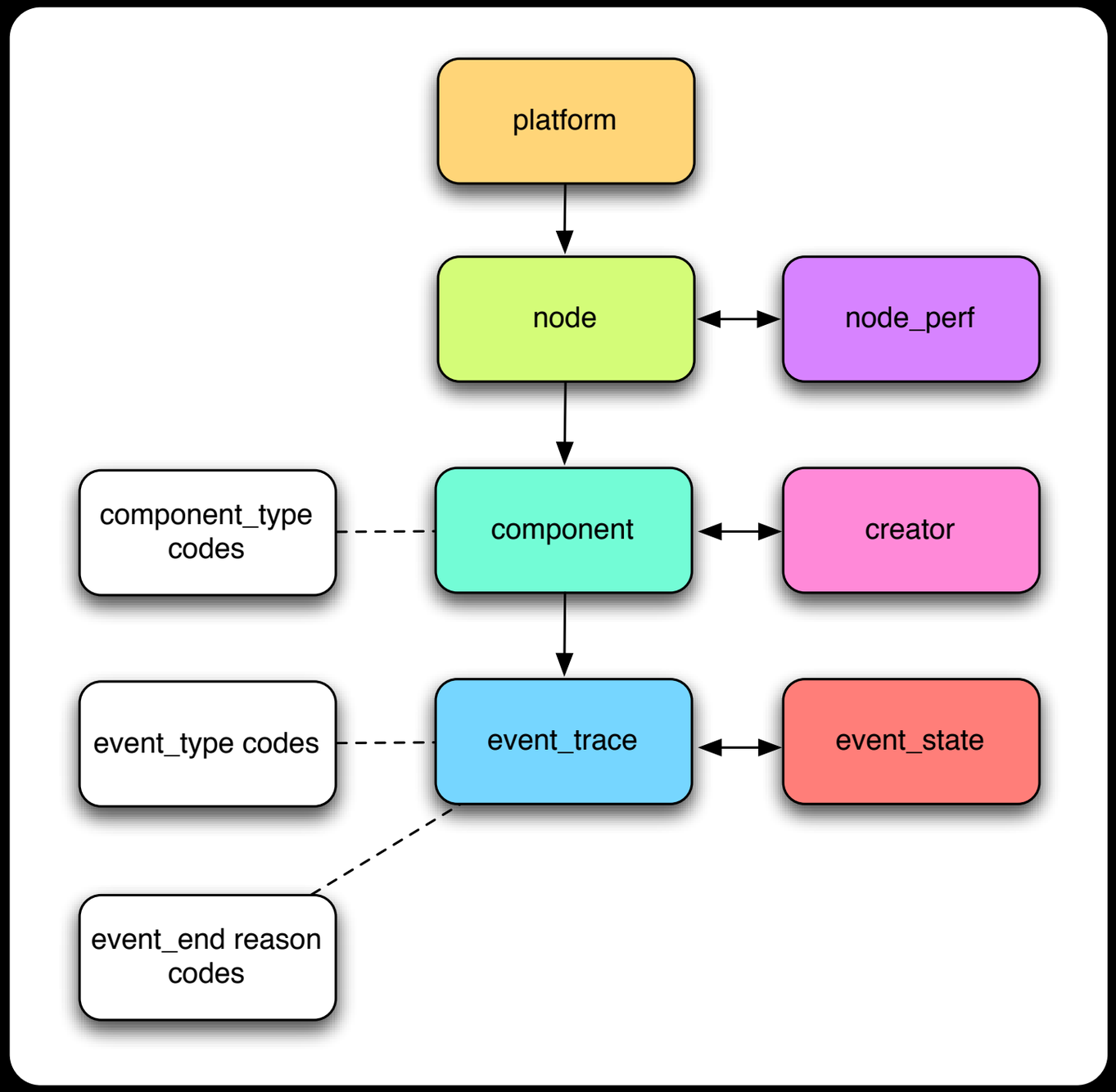
- Resource (versus job or user) centric
- Event-based
- Associated metadata
- Codes for different components, events, and errors

FTA Schema



- Resource (versus job or user) centric
- Event-based
- Associated metadata
- Codes for different components, events, and errors

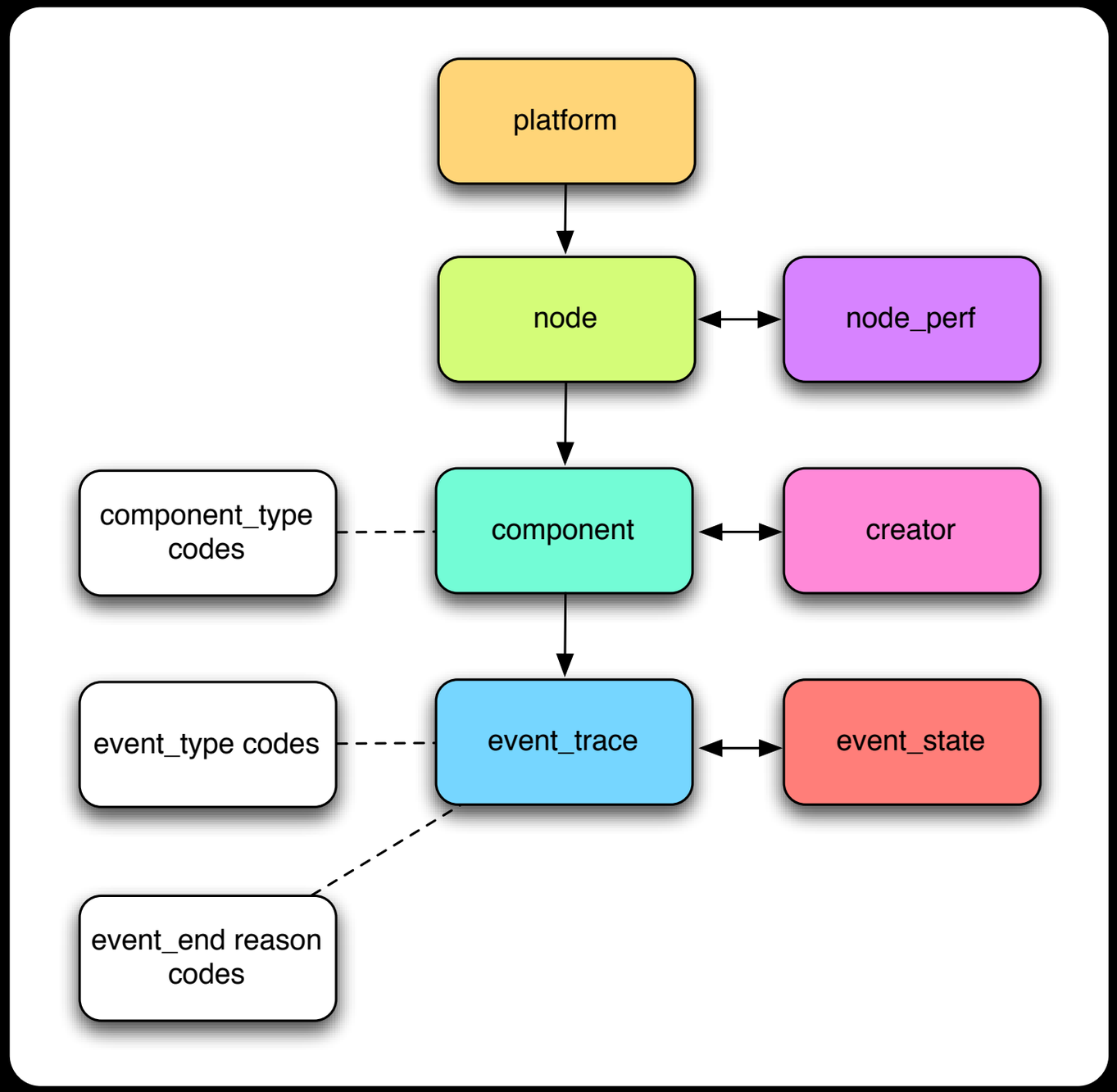
FTA Schema



- Resource (versus job or user) centric
- Event-based
- Associated metadata
- Codes for different components, events, and errors

- Balance between completeness and sparseness

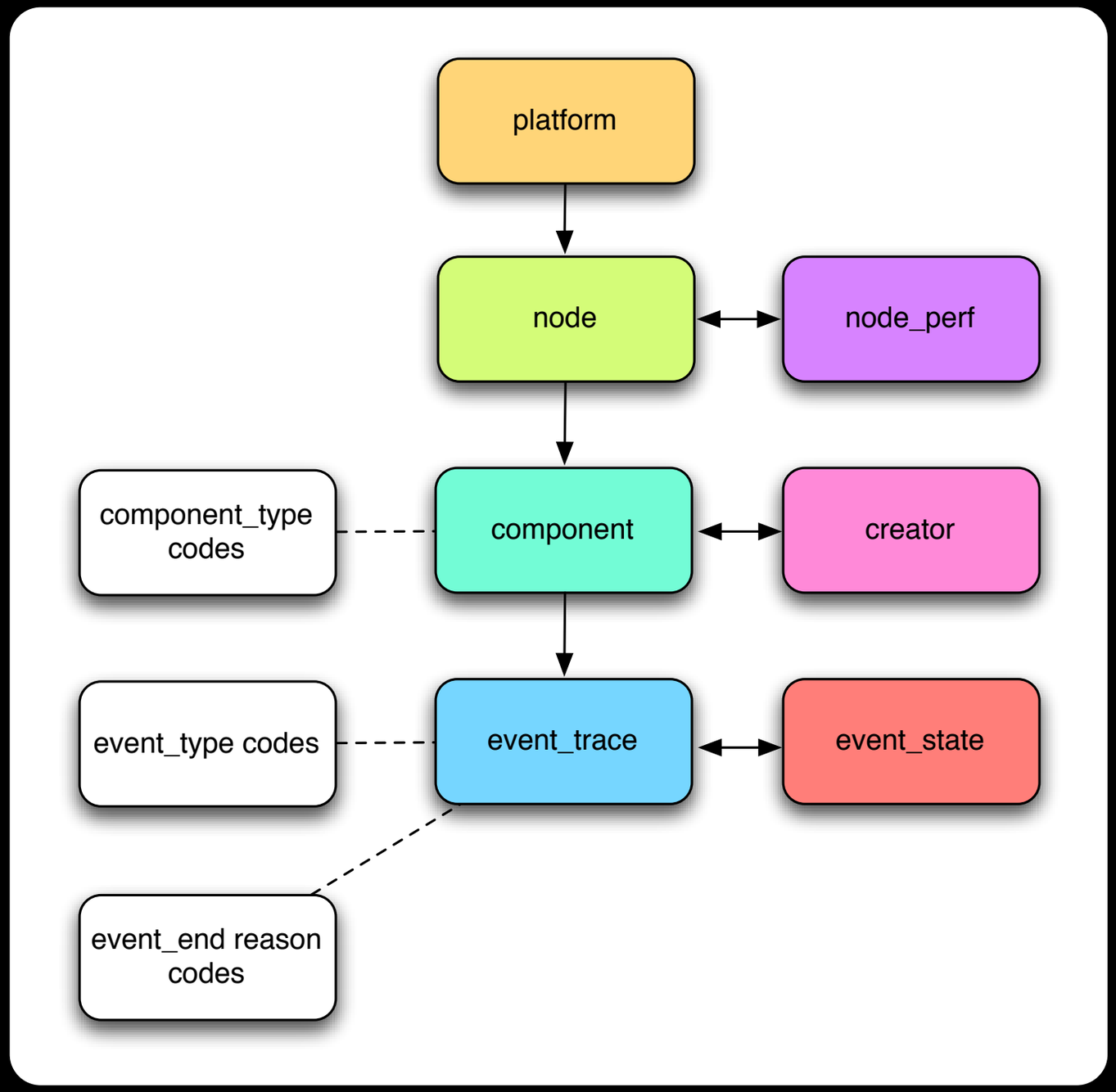
FTA Schema



- Resource (versus job or user) centric
- Event-based
- Associated metadata
- Codes for different components, events, and errors

- Balance between completeness and sparseness
- Extensibility

FTA Schema



- Raw, Tabbed, Relational database (MySQL)

- Resource (versus job or user) centric
- Event-based
- Associated metadata
- Codes for different components, events, and errors

- Balance between completeness and sparseness
- Extensibility

Data Quality Assessment

- Syntactic: standard format library that checks data types, number fields (automated)
- Semantic: time moves forward and is non-overlapping, state is valid (automated)
- Visual: look at the distribution for outliers (manual)

Data Sets

- Usage (p2p, supercomputer, grids, desktop PC's)
- Type (CPU, network, IO)
- Scale (50-240,000 hosts)
- Volatility (minutes to days)
- Resolution (wrt failure detection)

Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

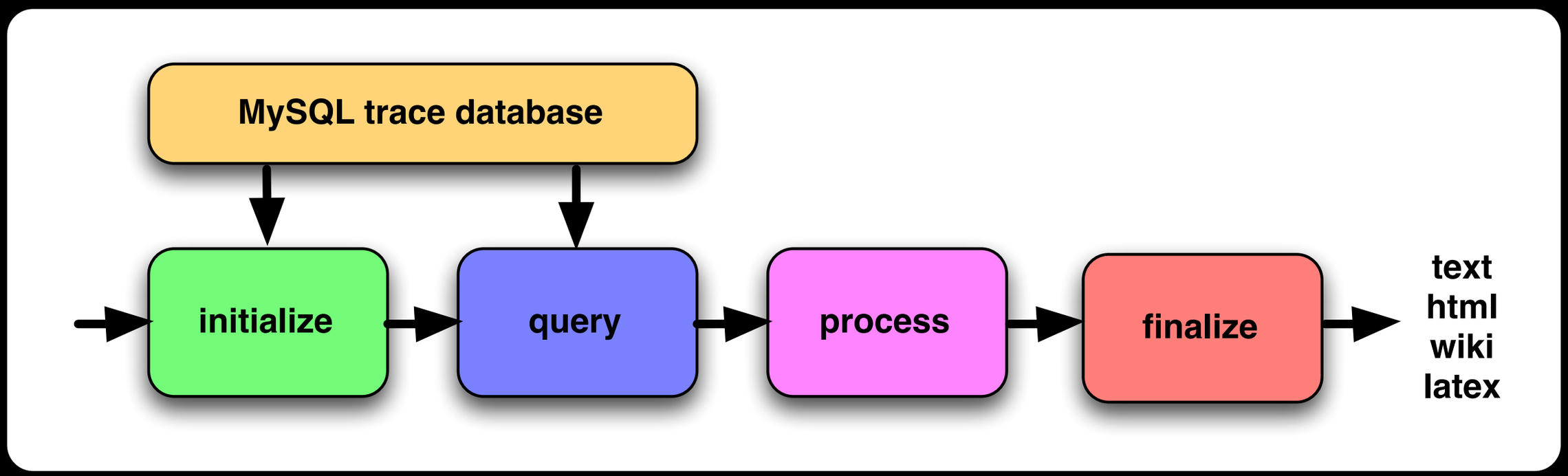
Currently 21 Data Sets

System	Type	# of Nodes	Target Component	Period	Year
SETI@home	Desktop Grid	226,208	CPU	1.5 years	2007-2009
Overnet	P2P	3,000	host	2 weeks	2003
Microsoft	Desktop	51,663	host	35 days	1999
LANL	SMP, HPC Clusters	4750	host	9 years	1996-2005
HPC2	HPC Clusters	256	IO	2.5 years	1996-2005
HPC4	Supercomputers	152516	every thing	~1 year	2004-2006
PNNL	HPC Cluster	980	host, network	4 years	2003-2007
NERSC	HPC Clusters	NA	IO	5 years	2001-2006
Skype	P2P	4000	host	1 month	2005
Web sites	Web servers	129	host	8 months	2001-2002
DNS	DNS servers	62,201	host	2 weeks	2004
PlanetLab	P2P	200-400	host	1.5 year	2004-2005
Grenouille03	DSL	4800	host	1 year	2003
Grenouille05	DSL	4800	host	1 year	2005
EGEE	Grid	2500 queues	CE queue	1 month	2007
Grid'5000	Grid	1288	host	1.5 years	2005-2006
Notre Dame	Desktop Grid	700	CPU, host	6 months	2007
ucb94	Desktop Grid	85	CPU	46 days	1994
sdsc03	Desktop Grid	275	CPU	1 month	2003
lri05	Desktop Grid	40	CPU	1 month	2005
deug05	Desktop Grid	40	CPU	1 month	2005

http://fta.inria.fr

Statistical Analysis

FTA Toolbox

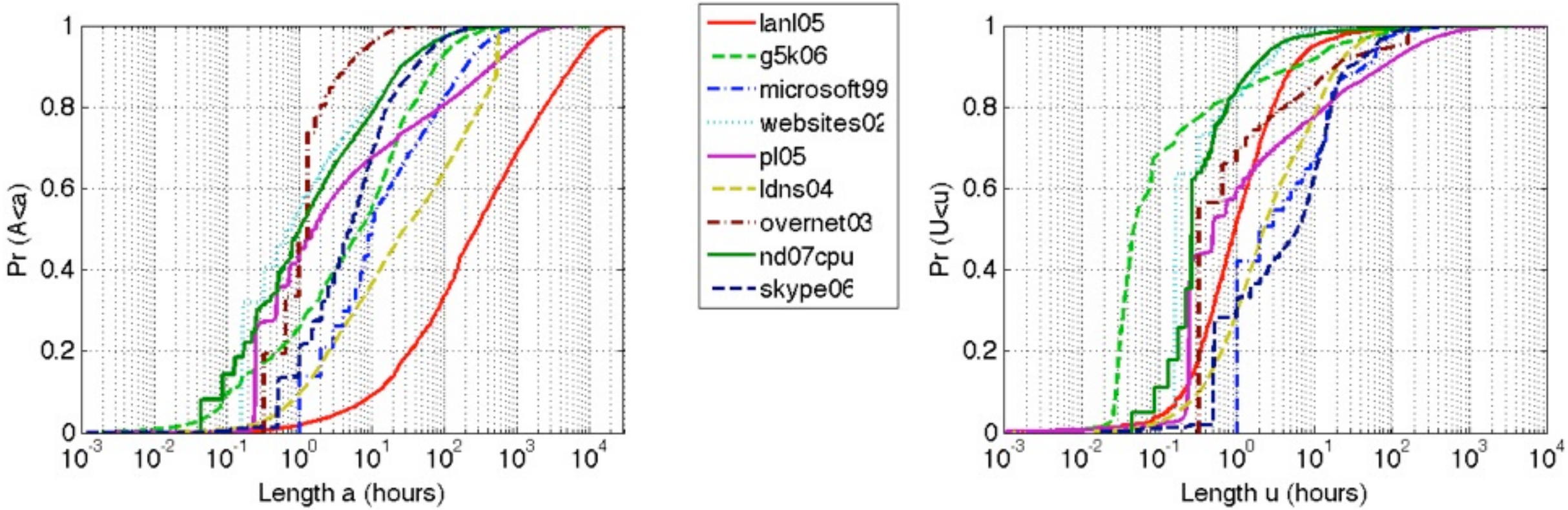


- Makes it easy to run a set of statistical measures across all the data sets
- Provides library of functions that can be reused and incorporated
- Implemented in Matlab
- svn checkout svn://scm.gforge.inria.fr/svn/fta/toolbox

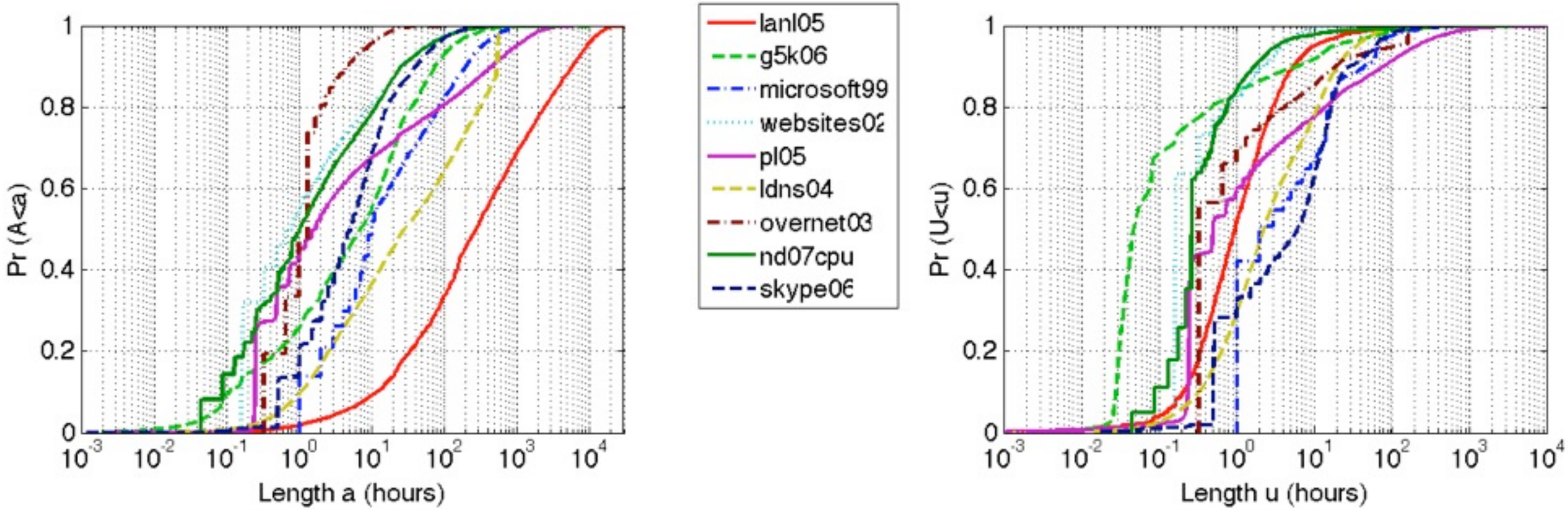
Failure Modelling

- Approach
  - Model availability and unavailability intervals, each with a single probability distribution
    - Assume availability and unavailability is identically and independently distributed
- Descriptive, not prescriptive

Distributions of Availability and Unavailability Intervals



Distributions of Availability and Unavailability Intervals



Qualitative Description

Parameter	Levels		Unit
	Low	High	
V: Volatility	$V < 50$	$V > 100$	hour
A: Availability	$A < 60$	$A > 90$	%
M: Measurement	$M < 6$	$M > 12$	month
S: Scale	$S < 1$	$S > 2$	10^3 nodes

Trace	V	A	M	S	Failure model
lanl05	L	H	H	H	Long-tailed/Long-tailed
g5k06	M	M	H	M	Long-tailed/Long-tailed
microsoft99	M	M	L	H	Short-tailed/Heavy-tailed
websites02	H	H	M	L	Long-tailed/Long-tailed
pl05	L	M	H	L	Long-tailed/Long-tailed
ldns04	L	H	L	H	Long-tailed/Short-tailed
overnet03	H	L	L	H	Short-tailed/Heavy-tailed
nd07cpu	H	M	M	L	Heavy-tailed/Long-tailed
skype06	H	L	L	H	Short-tailed/Short-tailed

Model Fitting

- For each candidate probability distribution
  - Compute parameters that maximize the distribution's likelihood
    - Measure goodness of fit using Kolomorov-Smirnov (KS) and Anderson-Darling (AD) tests
      - Compute p-value using 30 samples. Take average of 1000 p-values

P-Values for KS & AD Goodness-of-fit tests

Availability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.005 0.025	0.416 0.571	0.002 0.010	0.475 0.611	0.345 0.488
g5k06	0.012 0.038	0.472 0.597	0.003 0.018	0.394 0.564	0.409 0.507
microsoft99	0.005 0.084	0.294 0.546	0.000 0.049	0.371 0.611	0.198 0.418
websites02	0.000 0.006	0.079 0.354	0.000 0.027	0.188 0.401	0.055 0.182
pl05	0.000 0.000	0.080 0.245	0.002 0.016	0.168 0.321	0.043 0.131
ldns04	0.009 0.042	0.316 0.510	0.002 0.010	0.357 0.527	0.287 0.472
overnet03	0.045 0.460	0.068 0.532	0.000 0.013	0.160 0.660	0.052 0.481
nd07cpu	0.001 0.011	0.348 0.526	0.002 0.063	0.408 0.596	0.167 0.284
skype06	0.048 0.105	0.373 0.493	0.000 0.002	0.452 0.581	0.257 0.375

Unavailability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.000 0.004	0.196 0.346	0.000 0.001	0.481 0.607	0.042 0.095
g5k06	0.000 0.000	0.008 0.073	0.000 0.000	0.037 0.144	0.003 0.022
microsoft99	0.004 0.180	0.048 0.529	0.000 0.376	0.076 0.611	0.052 0.368
websites02	0.000 0.023	0.001 0.150	0.000 0.002	0.005 0.209	0.003 0.090
pl05	0.000 0.000	0.035 0.178	0.000 0.004	0.081 0.274	0.019 0.079
ldns04	0.035 0.112	0.404 0.538	0.000 0.001	0.464 0.607	0.277 0.411
overnet03	0.000 0.040	0.003 0.305	0.000 0.204	0.011 0.389	0.005 0.118
nd07cpu	0.000 0.004	0.028 0.219	0.000 0.031	0.126 0.559	0.003 0.032
skype06	0.071 0.191	0.288 0.478	0.002 0.015	0.182 0.449	0.267 0.408

P-Values for KS & AD Goodness-of-fit tests

p-value < 0.05 or 0.10
⇒ reject H0 that data came from fitted distribution

Availability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.005 0.025	0.416 0.571	0.002 0.010	0.475 0.611	0.345 0.488
g5k06	0.012 0.038	0.472 0.597	0.003 0.018	0.394 0.564	0.409 0.507
microsoft99	0.005 0.084	0.294 0.546	0.000 0.049	0.371 0.611	0.198 0.418
websites02	0.000 0.006	0.079 0.354	0.000 0.027	0.188 0.401	0.055 0.182
pl05	0.000 0.000	0.080 0.245	0.002 0.016	0.168 0.321	0.043 0.131
ldns04	0.009 0.042	0.316 0.510	0.002 0.010	0.357 0.527	0.287 0.472
overnet03	0.045 0.460	0.068 0.532	0.000 0.013	0.160 0.660	0.052 0.481
nd07cpu	0.001 0.011	0.348 0.526	0.002 0.063	0.408 0.596	0.167 0.284
skype06	0.048 0.105	0.373 0.493	0.000 0.002	0.452 0.581	0.257 0.375

Unavailability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.000 0.004	0.196 0.346	0.000 0.001	0.481 0.607	0.042 0.095
g5k06	0.000 0.000	0.008 0.073	0.000 0.000	0.037 0.144	0.003 0.022
microsoft99	0.004 0.180	0.048 0.529	0.000 0.376	0.076 0.611	0.052 0.368
websites02	0.000 0.023	0.001 0.150	0.000 0.002	0.005 0.209	0.003 0.090
pl05	0.000 0.000	0.035 0.178	0.000 0.004	0.081 0.274	0.019 0.079
ldns04	0.035 0.112	0.404 0.538	0.000 0.001	0.464 0.607	0.277 0.411
overnet03	0.000 0.040	0.003 0.305	0.000 0.204	0.011 0.389	0.005 0.118
nd07cpu	0.000 0.004	0.028 0.219	0.000 0.031	0.126 0.559	0.003 0.032
skype06	0.071 0.191	0.288 0.478	0.002 0.015	0.182 0.449	0.267 0.408

P-Values for KS & AD Goodness-of-fit tests

p-value < 0.05 or 0.10
⇒ reject H0 that data came from fitted distribution

Availability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.005 0.025	0.416 0.571	0.002 0.010	0.475 0.611	0.345 0.488
g5k06	0.012 0.038	0.472 0.597	0.003 0.018	0.394 0.564	0.409 0.507
microsoft99	0.005 0.084	0.294 0.546	0.000 0.049	0.371 0.611	0.198 0.418
websites02	0.000 0.006	0.079 0.354	0.000 0.027	0.188 0.401	0.055 0.182
pl05	0.000 0.000	0.080 0.245	0.002 0.016	0.168 0.321	0.043 0.131
ldns04	0.009 0.042	0.316 0.510	0.002 0.010	0.357 0.527	0.287 0.472
overnet03	0.045 0.460	0.068 0.532	0.000 0.013	0.160 0.660	0.052 0.481
nd07cpu	0.001 0.011	0.348 0.526	0.002 0.063	0.408 0.596	0.167 0.284
skype06	0.048 0.105	0.373 0.493	0.000 0.002	0.452 0.581	0.257 0.375

(Un)availability generally not heavy-tailed

Unavailability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.000 0.004	0.196 0.346	0.000 0.001	0.481 0.607	0.042 0.095
g5k06	0.000 0.000	0.008 0.073	0.000 0.000	0.037 0.144	0.003 0.022
microsoft99	0.004 0.180	0.048 0.529	0.000 0.376	0.076 0.611	0.052 0.368
websites02	0.000 0.023	0.001 0.150	0.000 0.002	0.005 0.209	0.003 0.090
pl05	0.000 0.000	0.035 0.178	0.000 0.004	0.081 0.274	0.019 0.079
ldns04	0.035 0.112	0.404 0.538	0.000 0.001	0.464 0.607	0.277 0.411
overnet03	0.000 0.040	0.003 0.305	0.000 0.204	0.011 0.389	0.005 0.118
nd07cpu	0.000 0.004	0.028 0.219	0.000 0.031	0.126 0.559	0.003 0.032
skype06	0.071 0.191	0.288 0.478	0.002 0.015	0.182 0.449	0.267 0.408

P-Values for KS & AD Goodness-of-fit tests

p-value < 0.05 or 0.10
⇒ reject H0 that data came from fitted distribution

Availability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.005 0.025	0.416 0.571	0.002 0.010	0.475 0.611	0.345 0.488
g5k06	0.012 0.038	0.472 0.597	0.003 0.018	0.394 0.564	0.409 0.507
microsoft99	0.005 0.084	0.294 0.546	0.000 0.049	0.371 0.611	0.198 0.418
websites02	0.000 0.006	0.079 0.354	0.000 0.027	0.188 0.401	0.055 0.182
pl05	0.000 0.000	0.080 0.245	0.002 0.016	0.168 0.321	0.043 0.131
ldns04	0.009 0.042	0.316 0.510	0.002 0.010	0.357 0.527	0.287 0.472
overnet03	0.045 0.460	0.068 0.532	0.000 0.013	0.160 0.660	0.052 0.481
nd07cpu	0.001 0.011	0.348 0.526	0.002 0.063	0.408 0.596	0.167 0.284
skype06	0.048 0.105	0.373 0.493	0.000 0.002	0.452 0.581	0.257 0.375

Unavailability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.000 0.004	0.196 0.346	0.000 0.001	0.481 0.607	0.042 0.095
g5k06	0.000 0.000	0.008 0.073	0.000 0.000	0.037 0.144	0.003 0.022
microsoft99	0.004 0.180	0.048 0.529	0.000 0.376	0.076 0.611	0.052 0.368
websites02	0.000 0.023	0.001 0.150	0.000 0.002	0.005 0.209	0.003 0.090
pl05	0.000 0.000	0.035 0.178	0.000 0.004	0.081 0.274	0.019 0.079
ldns04	0.035 0.112	0.404 0.538	0.000 0.001	0.464 0.607	0.277 0.411
overnet03	0.000 0.040	0.003 0.305	0.000 0.204	0.011 0.389	0.005 0.118
nd07cpu	0.000 0.004	0.028 0.219	0.000 0.031	0.126 0.559	0.003 0.032
skype06	0.071 0.191	0.288 0.478	0.002 0.015	0.182 0.449	0.267 0.408

P-Values for KS & AD Goodness-of-fit tests

p-value < 0.05 or 0.10
⇒ reject H0 that data came from fitted distribution

Availability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.005 0.025	0.416 0.571	0.002 0.010	0.475 0.611	0.345 0.488
g5k06	0.012 0.038	0.472 0.597	0.003 0.018	0.394 0.564	0.409 0.507
microsoft99	0.005 0.084	0.294 0.546	0.000 0.049	0.371 0.611	0.198 0.418
websites02	0.000 0.006	0.079 0.354	0.000 0.027	0.188 0.401	0.055 0.182
pl05	0.000 0.000	0.080 0.245	0.002 0.016	0.168 0.321	0.043 0.131
ldns04	0.009 0.042	0.316 0.510	0.002 0.010	0.357 0.527	0.287 0.472
overnet03	0.045 0.460	0.068 0.532	0.000 0.013	0.160 0.660	0.052 0.481
nd07cpu	0.001 0.011	0.348 0.526	0.002 0.063	0.408 0.596	0.167 0.284
skype06	0.048 0.105	0.373 0.493	0.000 0.002	0.452 0.581	0.257 0.375

Exponential usually not a good fit.

Unavailability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.000 0.004	0.196 0.346	0.000 0.001	0.481 0.607	0.042 0.095
g5k06	0.000 0.000	0.008 0.073	0.000 0.000	0.037 0.144	0.003 0.022
microsoft99	0.004 0.180	0.048 0.529	0.000 0.376	0.076 0.611	0.052 0.368
websites02	0.000 0.023	0.001 0.150	0.000 0.002	0.005 0.209	0.003 0.090
pl05	0.000 0.000	0.035 0.178	0.000 0.004	0.081 0.274	0.019 0.079
ldns04	0.035 0.112	0.404 0.538	0.000 0.001	0.464 0.607	0.277 0.411
overnet03	0.000 0.040	0.003 0.305	0.000 0.204	0.011 0.389	0.005 0.118
nd07cpu	0.000 0.004	0.028 0.219	0.000 0.031	0.126 0.559	0.003 0.032
skype06	0.071 0.191	0.288 0.478	0.002 0.015	0.182 0.449	0.267 0.408

P-Values for KS & AD Goodness-of-fit tests

p-value < 0.05 or 0.10
⇒ reject H0 that data came from fitted distribution

Availability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.005 0.025	0.416 0.571	0.002 0.010	0.475 0.611	0.345 0.488
g5k06	0.012 0.038	0.472 0.597	0.003 0.018	0.394 0.564	0.409 0.507
microsoft99	0.005 0.084	0.294 0.546	0.000 0.049	0.371 0.611	0.198 0.418
websites02	0.000 0.006	0.079 0.354	0.000 0.027	0.188 0.401	0.055 0.182
pl05	0.000 0.000	0.080 0.245	0.002 0.016	0.168 0.321	0.043 0.131
ldns04	0.009 0.042	0.316 0.510	0.002 0.010	0.357 0.527	0.287 0.472
overnet03	0.045 0.460	0.068 0.532	0.000 0.013	0.160 0.660	0.052 0.481
nd07cpu	0.001 0.011	0.348 0.526	0.002 0.063	0.408 0.596	0.167 0.284
skype06	0.048 0.105	0.373 0.493	0.000 0.002	0.452 0.581	0.257 0.375

Unavailability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.000 0.004	0.196 0.346	0.000 0.001	0.481 0.607	0.042 0.095
g5k06	0.000 0.000	0.008 0.073	0.000 0.000	0.037 0.144	0.003 0.022
microsoft99	0.004 0.180	0.048 0.529	0.000 0.376	0.076 0.611	0.052 0.368
websites02	0.000 0.023	0.001 0.150	0.000 0.002	0.005 0.209	0.003 0.090
pl05	0.000 0.000	0.035 0.178	0.000 0.004	0.081 0.274	0.019 0.079
ldns04	0.035 0.112	0.404 0.538	0.000 0.001	0.464 0.607	0.277 0.411
overnet03	0.000 0.040	0.003 0.305	0.000 0.204	0.011 0.389	0.005 0.118
nd07cpu	0.000 0.004	0.028 0.219	0.000 0.031	0.126 0.559	0.003 0.032
skype06	0.071 0.191	0.288 0.478	0.002 0.015	0.182 0.449	0.267 0.408

P-Values for KS & AD Goodness-of-fit tests

p-value < 0.05 or 0.10
⇒ reject H0 that data came from fitted distribution

Availability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.005 0.025	0.416 0.571	0.002 0.010	0.475 0.611	0.345 0.488
g5k06	0.012 0.038	0.472 0.597	0.003 0.018	0.394 0.564	0.409 0.507
microsoft99	0.005 0.084	0.294 0.546	0.000 0.049	0.371 0.611	0.198 0.418
websites02	0.000 0.006	0.079 0.354	0.000 0.027	0.188 0.401	0.055 0.182
pl05	0.000 0.000	0.080 0.245	0.002 0.016	0.168 0.321	0.043 0.131
ldns04	0.009 0.042	0.316 0.510	0.002 0.010	0.357 0.527	0.287 0.472
overnet03	0.045 0.460	0.068 0.532	0.000 0.013	0.160 0.660	0.052 0.481
nd07cpu	0.001 0.011	0.348 0.526	0.002 0.063	0.408 0.596	0.167 0.284
skype06	0.048 0.105	0.373 0.493	0.000 0.002	0.452 0.581	0.257 0.375

Unavailability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.000 0.004	0.196 0.346	0.000 0.001	0.481 0.607	0.042 0.095
g5k06	0.000 0.000	0.008 0.073	0.000 0.000	0.037 0.144	0.003 0.022
microsoft99	0.004 0.180	0.048 0.529	0.000 0.376	0.076 0.611	0.052 0.368
websites02	0.000 0.023	0.001 0.150	0.000 0.002	0.005 0.209	0.003 0.090
pl05	0.000 0.000	0.035 0.178	0.000 0.004	0.081 0.274	0.019 0.079
ldns04	0.035 0.112	0.404 0.538	0.000 0.001	0.464 0.607	0.277 0.411
overnet03	0.000 0.040	0.003 0.305	0.000 0.204	0.011 0.389	0.005 0.118
nd07cpu	0.000 0.004	0.028 0.219	0.000 0.031	0.126 0.559	0.003 0.032
skype06	0.071 0.191	0.288 0.478	0.002 0.015	0.182 0.449	0.267 0.408

Gamma a good fit. Amenable for Markov Models

P-Values for KS & AD Goodness-of-fit tests

p-value < 0.05 or 0.10
⇒ reject H0 that data came from fitted distribution

Availability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.005 0.025	0.416 0.571	0.002 0.010	0.475 0.611	0.345 0.488
g5k06	0.012 0.038	0.472 0.597	0.003 0.018	0.394 0.564	0.409 0.507
microsoft99	0.005 0.084	0.294 0.546	0.000 0.049	0.371 0.611	0.198 0.418
websites02	0.000 0.006	0.079 0.354	0.000 0.027	0.188 0.401	0.055 0.182
pl05	0.000 0.000	0.080 0.245	0.002 0.016	0.168 0.321	0.043 0.131
ldns04	0.009 0.042	0.316 0.510	0.002 0.010	0.357 0.527	0.287 0.472
overnet03	0.045 0.460	0.068 0.532	0.000 0.013	0.160 0.660	0.052 0.481
nd07cpu	0.001 0.011	0.348 0.526	0.002 0.063	0.408 0.596	0.167 0.284
skype06	0.048 0.105	0.373 0.493	0.000 0.002	0.452 0.581	0.257 0.375

Unavailability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.000 0.004	0.196 0.346	0.000 0.001	0.481 0.607	0.042 0.095
g5k06	0.000 0.000	0.008 0.073	0.000 0.000	0.037 0.144	0.003 0.022
microsoft99	0.004 0.180	0.048 0.529	0.000 0.376	0.076 0.611	0.052 0.368
websites02	0.000 0.023	0.001 0.150	0.000 0.002	0.005 0.209	0.003 0.090
pl05	0.000 0.000	0.035 0.178	0.000 0.004	0.081 0.274	0.019 0.079
ldns04	0.035 0.112	0.404 0.538	0.000 0.001	0.464 0.607	0.277 0.411
overnet03	0.000 0.040	0.003 0.305	0.000 0.204	0.011 0.389	0.005 0.118
nd07cpu	0.000 0.004	0.028 0.219	0.000 0.031	0.126 0.559	0.003 0.032
skype06	0.071 0.191	0.288 0.478	0.002 0.015	0.182 0.449	0.267 0.408

P-Values for KS & AD Goodness-of-fit tests

p-value < 0.05 or 0.10
⇒ reject H0 that data came from fitted distribution

Availability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.005 0.025	0.416 0.571	0.002 0.010	0.475 0.611	0.345 0.488
g5k06	0.012 0.038	0.472 0.597	0.003 0.018	0.394 0.564	0.409 0.507
microsoft99	0.005 0.084	0.294 0.546	0.000 0.049	0.371 0.611	0.198 0.418
websites02	0.000 0.006	0.079 0.354	0.000 0.027	0.188 0.401	0.055 0.182
pl05	0.000 0.000	0.080 0.245	0.002 0.016	0.168 0.321	0.043 0.131
ldns04	0.009 0.042	0.316 0.510	0.002 0.010	0.357 0.527	0.287 0.472
overnet03	0.045 0.460	0.068 0.532	0.000 0.013	0.160 0.660	0.052 0.481
nd07cpu	0.001 0.011	0.348 0.526	0.002 0.063	0.408 0.596	0.167 0.284
skype06	0.048 0.105	0.373 0.493	0.000 0.002	0.452 0.581	0.257 0.375

Unavailability

Trace	Exponential	Weibull	Pareto	Log-Normal	Gamma
lanl05	0.000 0.004	0.196 0.346	0.000 0.001	0.481 0.607	0.042 0.095
g5k06	0.000 0.000	0.008 0.073	0.000 0.000	0.037 0.144	0.003 0.022
microsoft99	0.004 0.180	0.048 0.529	0.000 0.376	0.076 0.611	0.052 0.368
websites02	0.000 0.023	0.001 0.150	0.000 0.002	0.005 0.209	0.003 0.090
pl05	0.000 0.000	0.035 0.178	0.000 0.004	0.081 0.274	0.019 0.079
ldns04	0.035 0.112	0.404 0.538	0.000 0.001	0.464 0.607	0.277 0.411
overnet03	0.000 0.040	0.003 0.305	0.000 0.204	0.011 0.389	0.005 0.118
nd07cpu	0.000 0.004	0.028 0.219	0.000 0.031	0.126 0.559	0.003 0.032
skype06	0.071 0.191	0.288 0.478	0.002 0.015	0.182 0.449	0.267 0.408

Weibull and Log Normal provide best fit

Parameters of Distributions

	Availability				Unavailability			
Trace	**Exp**(μ)	**Wbl**(k, λ)	**LogN**(μ, σ)	**Gam**(k, λ)	**Exp**(μ)	**Wbl**(k, λ)	**LogN**(μ, σ)	**Gam**(k, λ)
lanl05	1779.99	0.48 816.60	5.56 2.39	0.35 5102.71	5.92	0.58 2.18	0.05 1.42	0.38 15.44
g5k06	32.41	0.48 14.37	1.51 2.42	0.34 94.35	7.41	0.35 0.47	-2.00 2.20	0.19 39.92
microsoft99	67.01	0.55 35.30	2.62 1.84	0.41 162.19	16.49	0.60 9.34	1.42 1.54	0.46 35.52
websites02	11.85	0.46 3.68	0.23 2.02	0.31 38.67	1.18	0.65 0.61	-1.12 1.13	0.50 2.37
pl05	159.49	0.33 19.35	1.44 2.86	0.20 788.03	49.61	0.36 5.59	0.40 2.45	0.21 237.65
ldns04	141.06	0.51 79.30	3.25 2.33	0.39 362.43	8.61	0.63 5.62	0.91 1.64	0.51 16.87
overnet03	2.29	0.85 2.04	0.19 0.98	0.91 2.53	12.00	0.44 2.98	0.08 1.80	0.29 41.64
nd07cpu	13.73	0.45 4.16	0.30 2.20	0.30 46.16	4.25	0.51 0.74	-1.02 1.27	0.28 15.07
skype06	16.27	0.64 10.86	1.60 1.57	0.53 30.79	14.31	0.63 9.48	1.40 1.73	0.50 28.53

μ: mean, σ: std dev.,
k: shape, λ: scale

Parameters of Distributions

Availability

Unavailability

Trace	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)
lanl05	1779.99	0.48 816.60	5.56 2.39	0.35 5102.71	5.92	0.58 2.18	0.05 1.42	0.38 15.44
g5k06	32.41	0.48 14.37	1.51 2.42	0.34 94.35	7.41	0.35 0.47	-2.00 2.20	0.19 39.92
microsoft99	67.01	0.55 35.30	2.62 1.84	0.41 162.19	16.49	0.60 9.34	1.42 1.54	0.46 35.52
websites02	11.85	0.46 3.68	0.23 2.02	0.31 38.67	1.18	0.65 0.61	-1.12 1.13	0.50 2.37
pl05	159.49	0.33 19.35	1.44 2.86	0.20 788.03	49.61	0.36 5.59	0.40 2.45	0.21 237.65
ldns04	141.06	0.51 79.30	3.25 2.33	0.39 362.43	8.61	0.63 5.62	0.91 1.64	0.51 16.87
overnet03	2.29	0.85 2.04	0.19 0.98	0.91 2.53	12.00	0.44 2.98	0.08 1.80	0.29 41.64
nd07cpu	13.73	0.45 4.16	0.30 2.20	0.30 46.16	4.25	0.51 0.74	-1.02 1.27	0.28 15.07
skype06	16.27	0.64 10.86	1.60 1.57	0.53 30.79	14.31	0.63 9.48	1.40 1.73	0.50 28.53

μ: mean, σ: std dev.,
k: shape, λ: scale

$k < 1$,
∴ decreasing hazard rate

Can different interpretations of trace data sets affect the model?

Ambiguous Data Sets

Data Set	Ambiguity	Interpretation
G5K06	Monitored state is an error or failure	error
G5K06B		failure

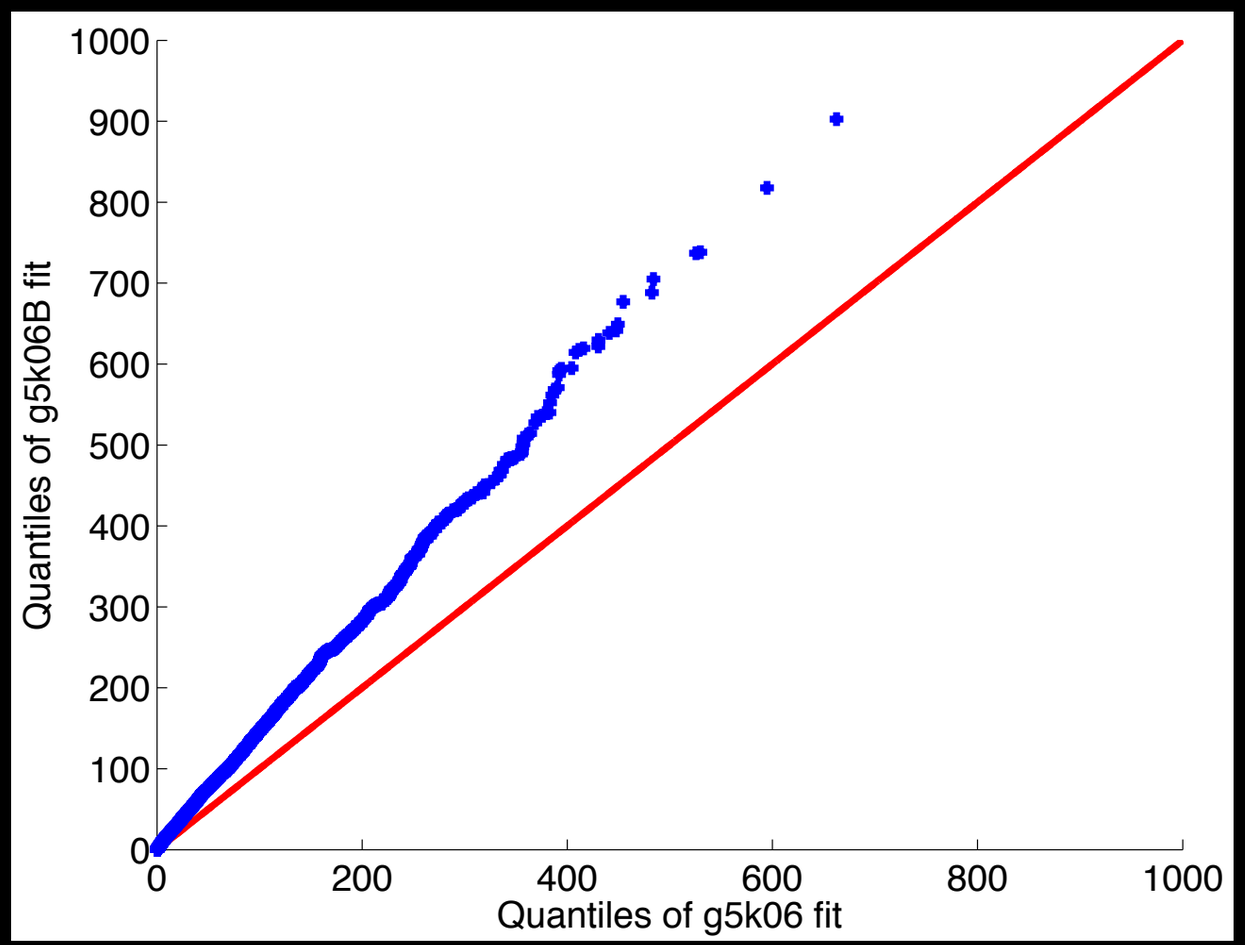
Ambiguous Data Sets

Data Set	Ambiguity	Interpretation
G5K06	Monitored state is an error or failure	error
G5K06B		failure
LANL0516	Overlapping intervals	union
LANL0516B		intersection

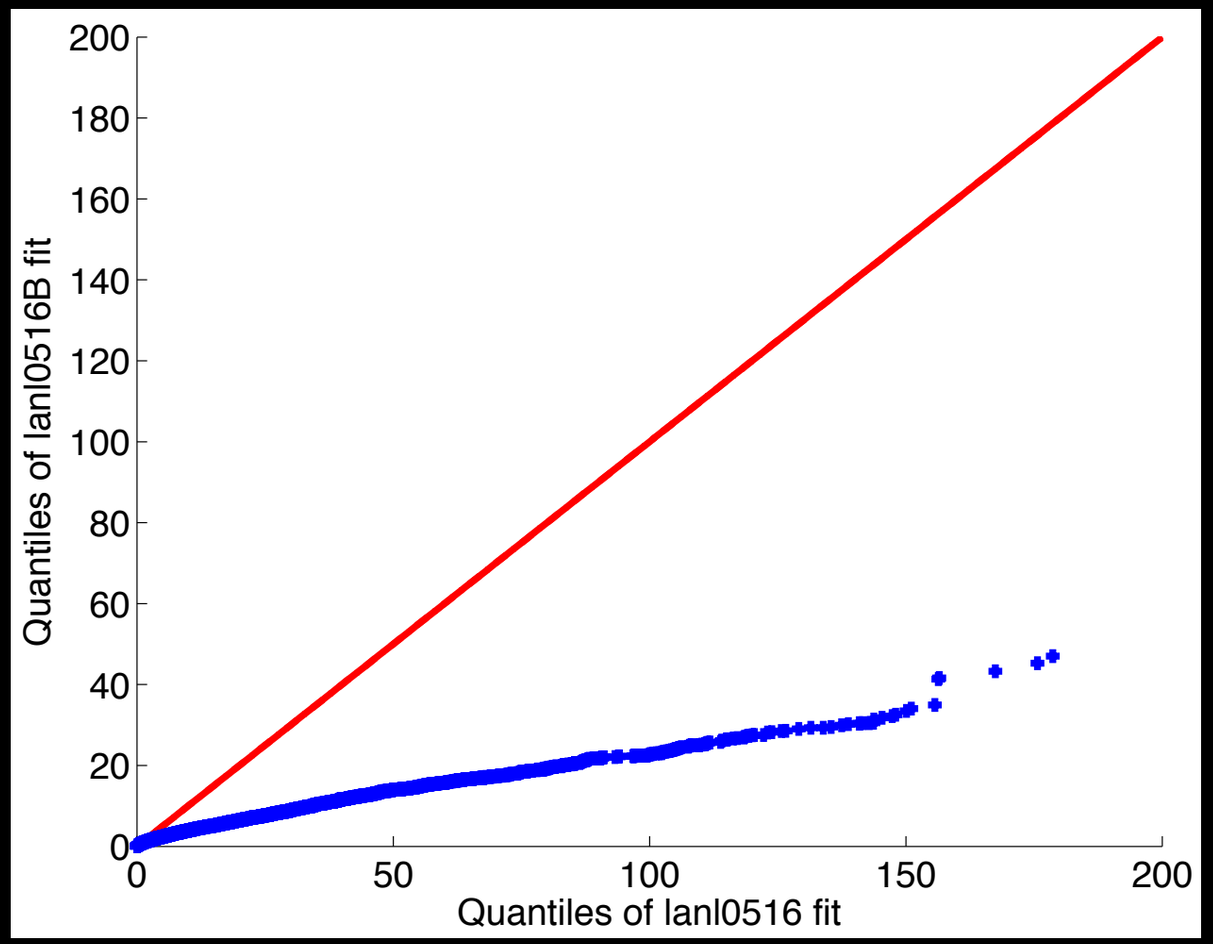
Ambiguous Data Sets

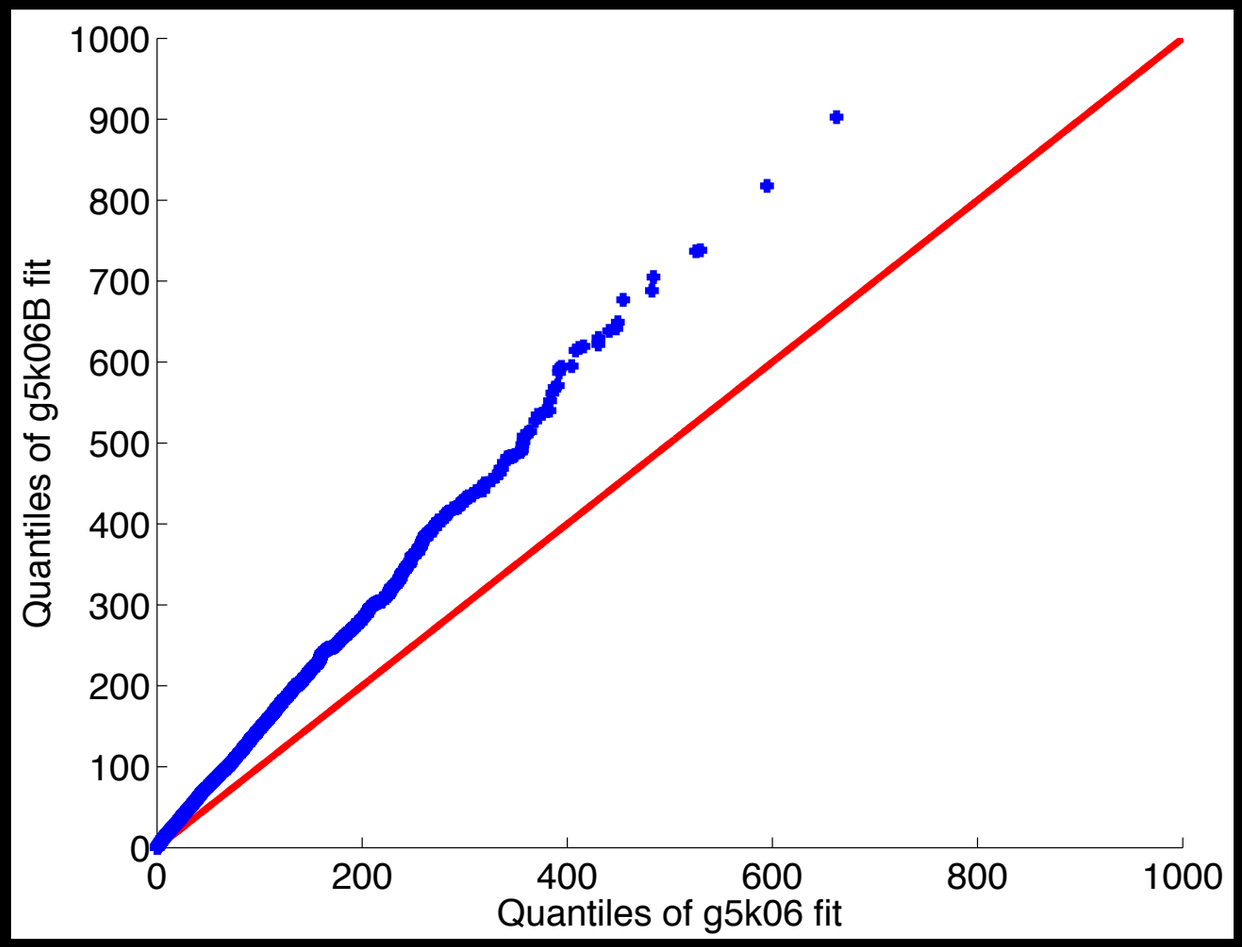
Data Set	Ambiguity	Interpretation
G5K06	Monitored state is an error or failure	error
G5K06B		failure
LANL0516	Overlapping intervals	union
LANL0516B		intersection
ND07CPU	Definition of idleness	w/o user and CPU load for 15 mins
ND07CPUB		CPU load < 10%

QQ Plots for Ambiguous Data Sets

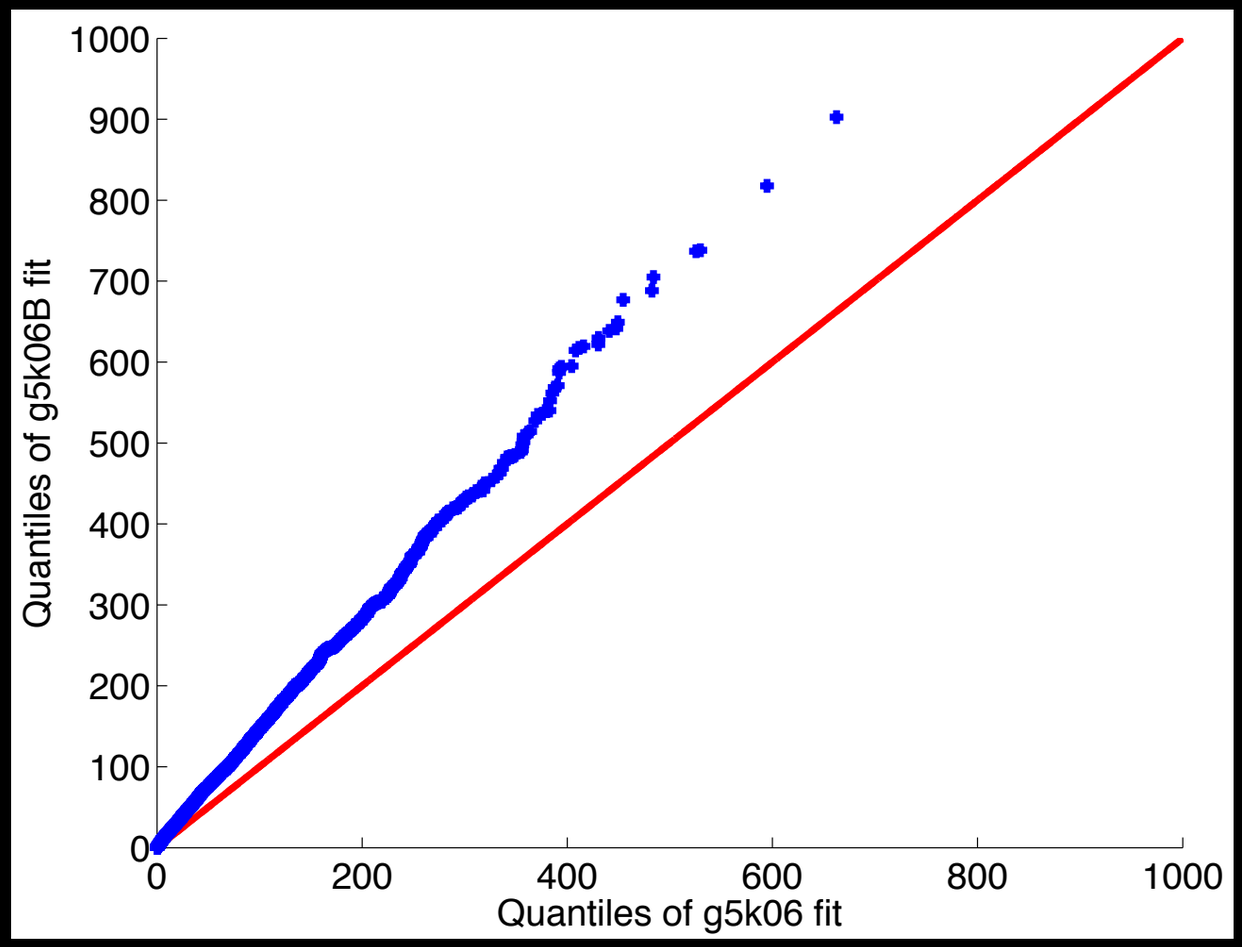


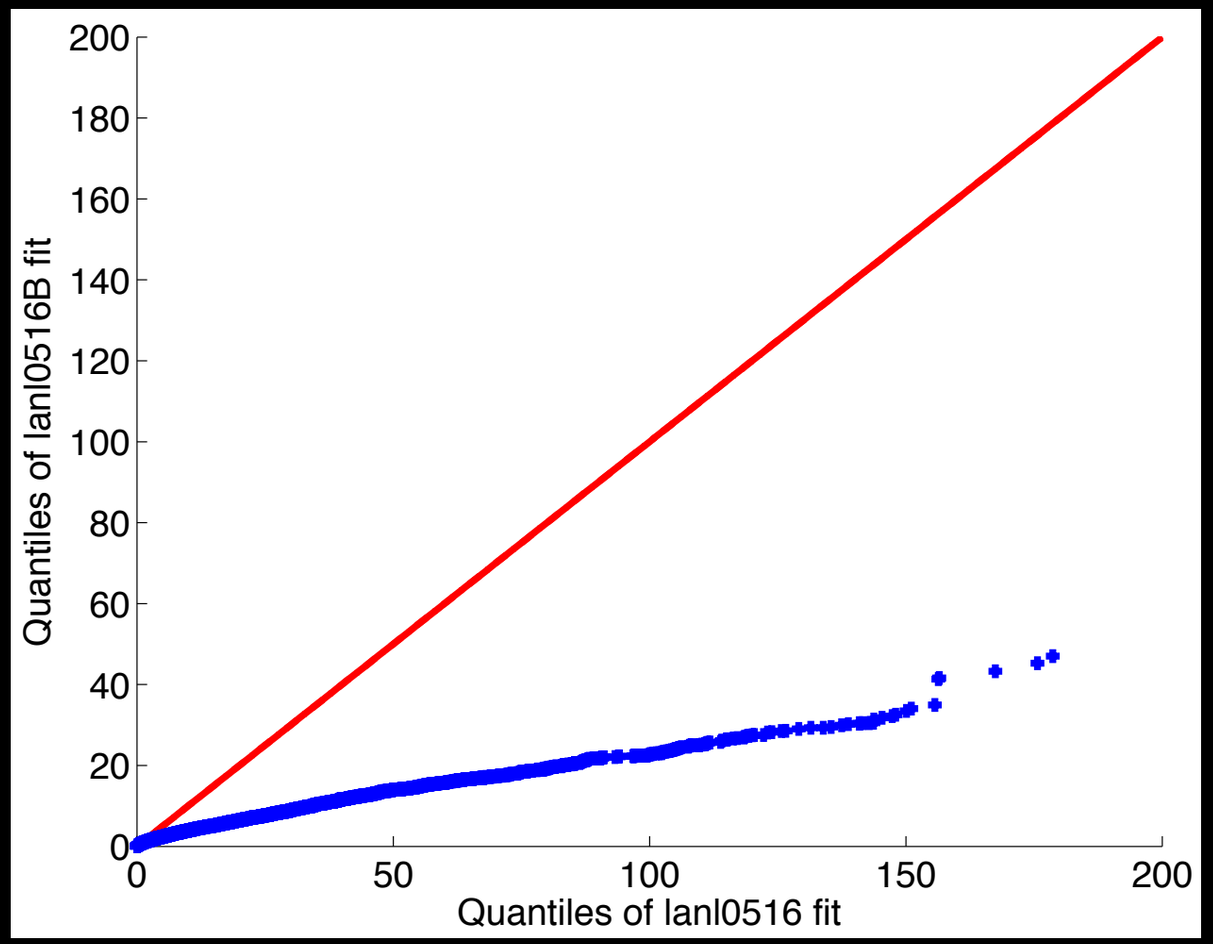
QQ Plots for Ambiguous Data Sets

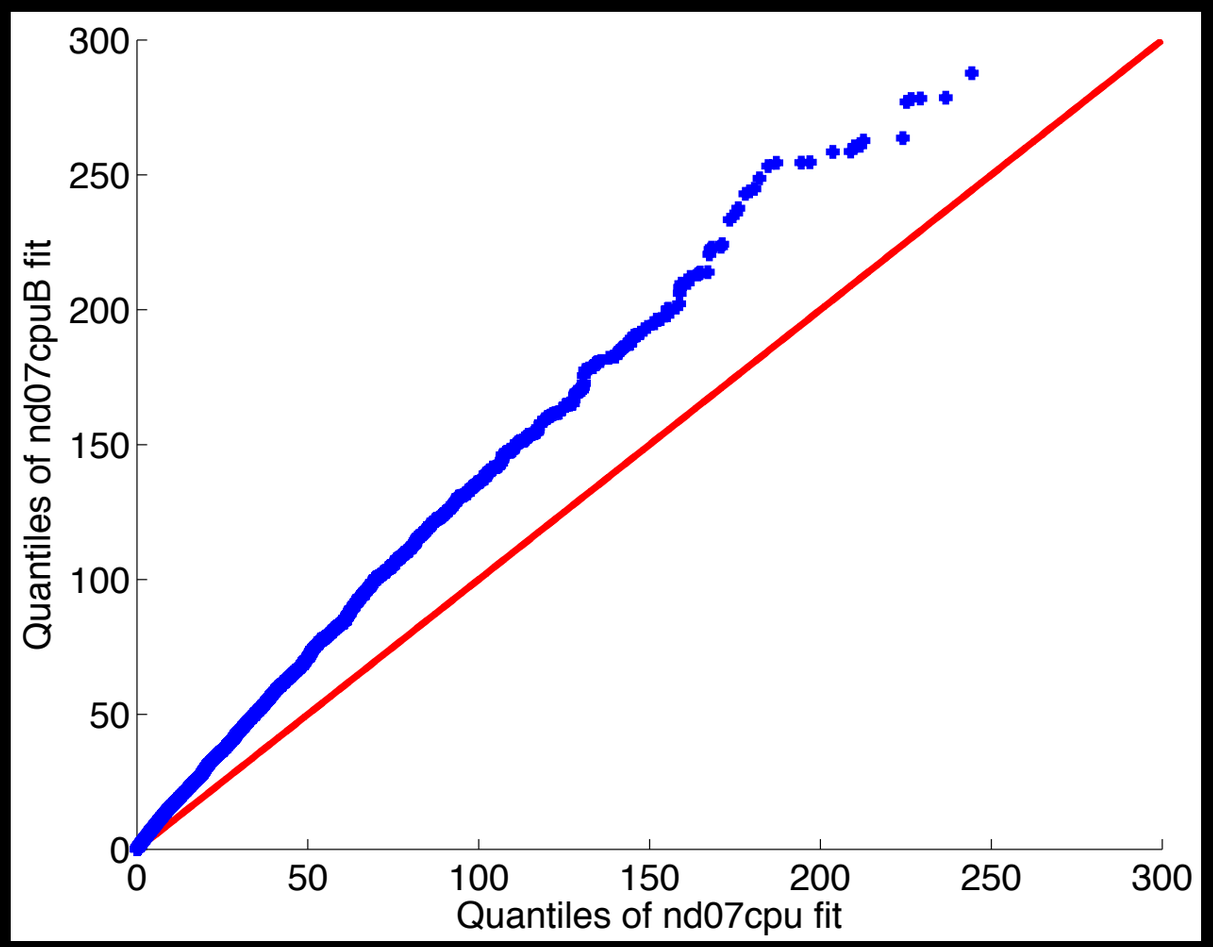




QQ Plots for Ambiguous Data Sets







Distribution Parameters for Ambiguous Data Sets

System	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)
g5k06	32.41	0.48 14.37	1.51 2.42	0.34 94.35	7.41	0.35 0.47	-2.00 2.20	0.19 39.92
g5k06B	48.61	0.52 22.66	2.08 2.21	0.37 131.78	6.54	0.35 0.31	-2.36 2.07	0.18 37.00
lanl05	1779.99	0.48 816.60	5.56 2.39	0.35 5102.71	5.92	0.58 2.18	0.05 1.42	0.38 15.44
lanl05B	1774.21	0.48 812.98	5.55 2.39	0.35 5087.60	5.06	0.59 2.12	0.03 1.40	0.41 12.28
nd07cpu	13.73	0.45 4.16	0.30 2.20	0.30 46.16	4.25	0.51 0.74	-1.02 1.27	0.28 15.07
nd07cpuB	20.12	0.48 7.21	0.91 2.07	0.33 61.74	5.75	0.49 0.83	-0.91 1.21	0.26 21.72

μ: mean, σ: std dev.,
k: shape, λ: scale

Distribution Parameters for Ambiguous Data Sets

Mean of G5K06B 1.5 times greater than G5K06

System	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)
g5k06	32.41	0.48 14.37	1.51 2.42	0.34 94.35	7.41	0.35 0.47	-2.00 2.20	0.19 39.92
g5k06B	48.61	0.52 22.66	2.08 2.21	0.37 131.78	6.54	0.35 0.31	-2.36 2.07	0.18 37.00
lanl05	1779.99	0.48 816.60	5.56 2.39	0.35 5102.71	5.92	0.58 2.18	0.05 1.42	0.38 15.44
lanl05B	1774.21	0.48 812.98	5.55 2.39	0.35 5087.60	5.06	0.59 2.12	0.03 1.40	0.41 12.28
nd07cpu	13.73	0.45 4.16	0.30 2.20	0.30 46.16	4.25	0.51 0.74	-1.02 1.27	0.28 15.07
nd07cpuB	20.12	0.48 7.21	0.91 2.07	0.33 61.74	5.75	0.49 0.83	-0.91 1.21	0.26 21.72

μ: mean, σ: std dev.,
k: shape, λ: scale

Distribution Parameters for Ambiguous Data Sets

System	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)
g5k06	32.41	0.48 14.37	1.51 2.42	0.34 94.35	7.41	0.35 0.47	-2.00 2.20	0.19 39.92
g5k06B	48.61	0.52 22.66	2.08 2.21	0.37 131.78	6.54	0.35 0.31	-2.36 2.07	0.18 37.00
lanl05	1779.99	0.48 816.60	5.56 2.39	0.35 5102.71	5.92	0.58 2.18	0.05 1.42	0.38 15.44
lanl05B	1774.21	0.48 812.98	5.55 2.39	0.35 5087.60	5.06	0.59 2.12	0.03 1.40	0.41 12.28
nd07cpu	13.73	0.45 4.16	0.30 2.20	0.30 46.16	4.25	0.51 0.74	-1.02 1.27	0.28 15.07
nd07cpuB	20.12	0.48 7.21	0.91 2.07	0.33 61.74	5.75	0.49 0.83	-0.91 1.21	0.26 21.72

μ: mean, σ: std dev.,
k: shape, λ: scale

Distribution Parameters for Ambiguous Data Sets

System	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)	Exp(μ)	Wbl(k, λ)	LogN(μ, σ)	Gam(k, λ)
g5k06	32.41	0.48 14.37	1.51 2.42	0.34 94.35	7.41	0.35 0.47	-2.00 2.20	0.19 39.92
g5k06B	48.61	0.52 22.66	2.08 2.21	0.37 131.78	6.54	0.35 0.31	-2.36 2.07	0.18 37.00
lanl05	1779.99	0.48 816.60	5.56 2.39	0.35 5102.71	5.92	0.58 2.18	0.05 1.42	0.38 15.44
lanl05B	1774.21	0.48 812.98	5.55 2.39	0.35 5087.60	5.06	0.59 2.12	0.03 1.40	0.41 12.28
nd07cpu	13.73	0.45 4.16	0.30 2.20	0.30 46.16	4.25	0.51 0.74	-1.02 1.27	0.28 15.07
nd07cpuB	20.12	0.48 7.21	0.91 2.07	0.33 61.74	5.75	0.49 0.83	-0.91 1.21	0.26 21.72

Gamma scale parameter often significantly different

μ: mean, σ: std dev.,
k: shape, λ: scale

How to identify interpretation?

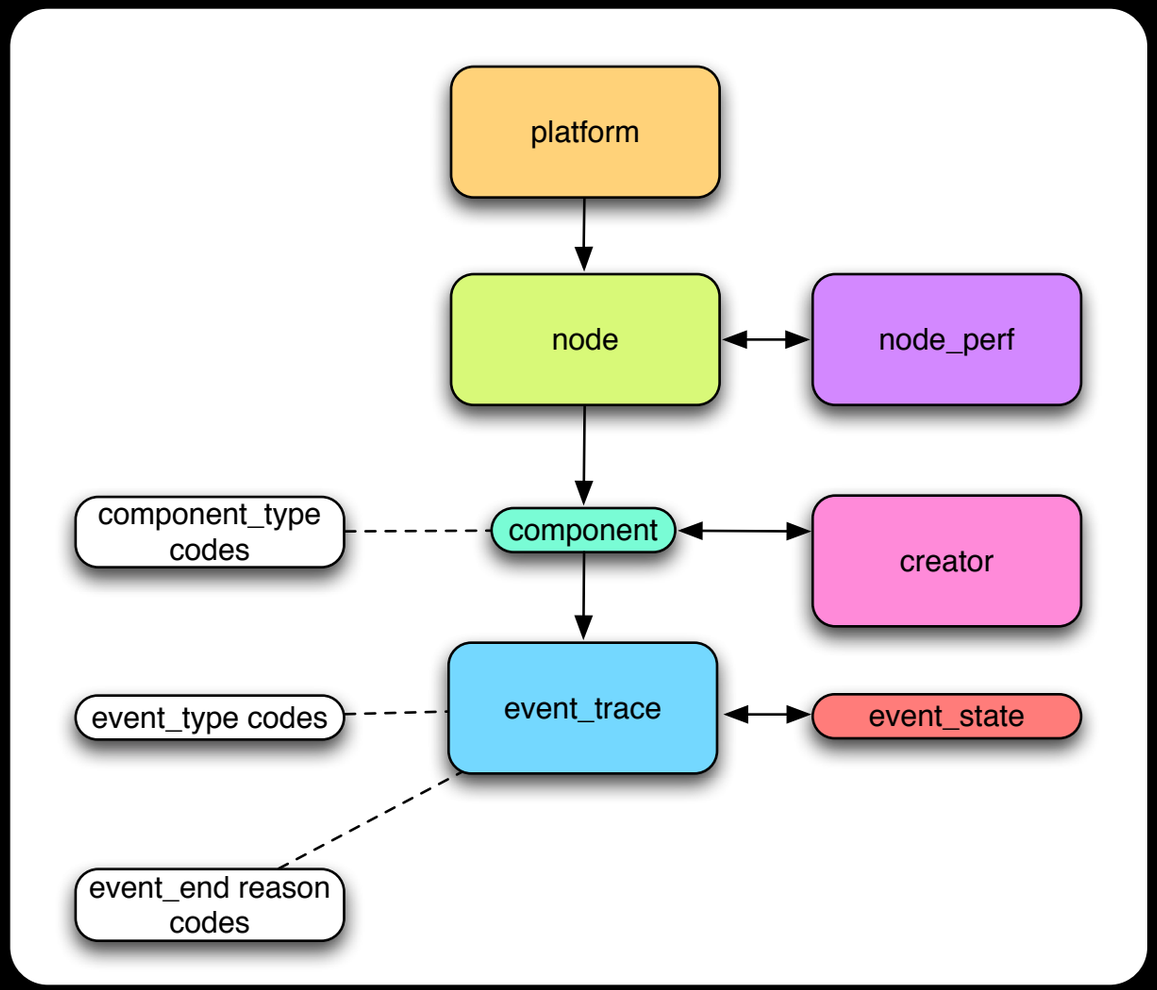
- Parsing script is the exact interpretation
  - Meaning explained in comments
    - Publicly accessible in svn
- Format supports different interpretations of availability
  - Can have multiple event_trace's corresponding to different definitions availability
- So each interpretation can be uniquely identified

How to resolve differences of interpretation?

- Determine which interpretation affects the application. (E.g. G5K06)
- Determine most common interpretation, or interpretation that is the lowest common denominator (E.g. ND07CPU)
- Exclude period of ambiguity or post-process it so that it is consistent with rest of data set (E.g. LANL05)

Future Directions

- Call to arms: trace data exists in many production environments, but not always accessible
  - Include more production systems
    - Types of failures
    - Causes of failures
    - State before failures
    - Automated trace collection
    - Failure models and algorithms
    - Integration of job and resource failures



Acknowledgements

- All contributors of trace data to the FTA
- INRIA ALEAE project directed by Emmanuel Jeannot
- Feedback from Cecile Germain, Eric Heien, Artur Andrzejak, anonymous reviewers

Summary

- FTA: Data Sets, Format, Tools
  - http://fta.inria.fr
- High-level modelling and statistical characterization of 9 data sets
- Slight differences in interpretation make significant difference in model
- Got data? Questions? Please email derrick.kondo@inria.fr or any other FTA team member

Thank you