A Model for space-correlated Failures in Large-Scale Distributed Systems

Matthieu Gallet

joint work with Nezih Yigitbasi, Bahman Javadi, Derrick Kondo, Alexandru Iosup, Dick Epema
CNRS
INRIA GRAAL project-team
Laboratoire de l'Informatique du Parallélisme
École Normale Supérieure de Lyon, France

ALEAE, March 15, 2009

Introduction

- Distributed systems more and more widespread
- Grids, peer-to-peer file sharing, distributed computation networks,
- Large scale and complexity of platforms
- Nodes may leave the system, join it or simply fail

Introduction

- Distributed systems more and more widespread
- Grids, peer-to-peer file sharing, distributed computation networks,
- Large scale and complexity of platforms
- Nodes may leave the system, join it or simply fail

Burstiness of failures

- Need to understand the characteristics of failures
- Bursty behaviour of failures: many failures occur within a short period
- VAX cluster: 58% of failures occured in bursts
- Grid'5000: 30% of failures involve multiple machines
- Bursty arrivals break assumptions made by fault tolerant algorithms
- Called *space-correlated* failures

Burstiness of failures

- Need to understand the characteristics of failures
- Bursty behaviour of failures: many failures occur within a short period
- VAX cluster: 58% of failures occured in bursts
- Grid'5000: 30% of failures involve multiple machines
- Bursty arrivals break assumptions made by fault tolerant algorithms
- Called *space-correlated* failures

Burstiness of failures

- Need to understand the characteristics of failures
- Bursty behaviour of failures: many failures occur within a short period
- VAX cluster: 58% of failures occured in bursts
- Grid'5000: 30% of failures involve multiple machines
- Bursty arrivals break assumptions made by fault tolerant algorithms
- Called *space-correlated* failures

Outline

- Material: as many real traces as possible
- Problem definition
- Model components
- Result analysis

Failure Trace Archive

- Joint project of Delft University, INRIA, Osaka University, Zuse Institute Berlin

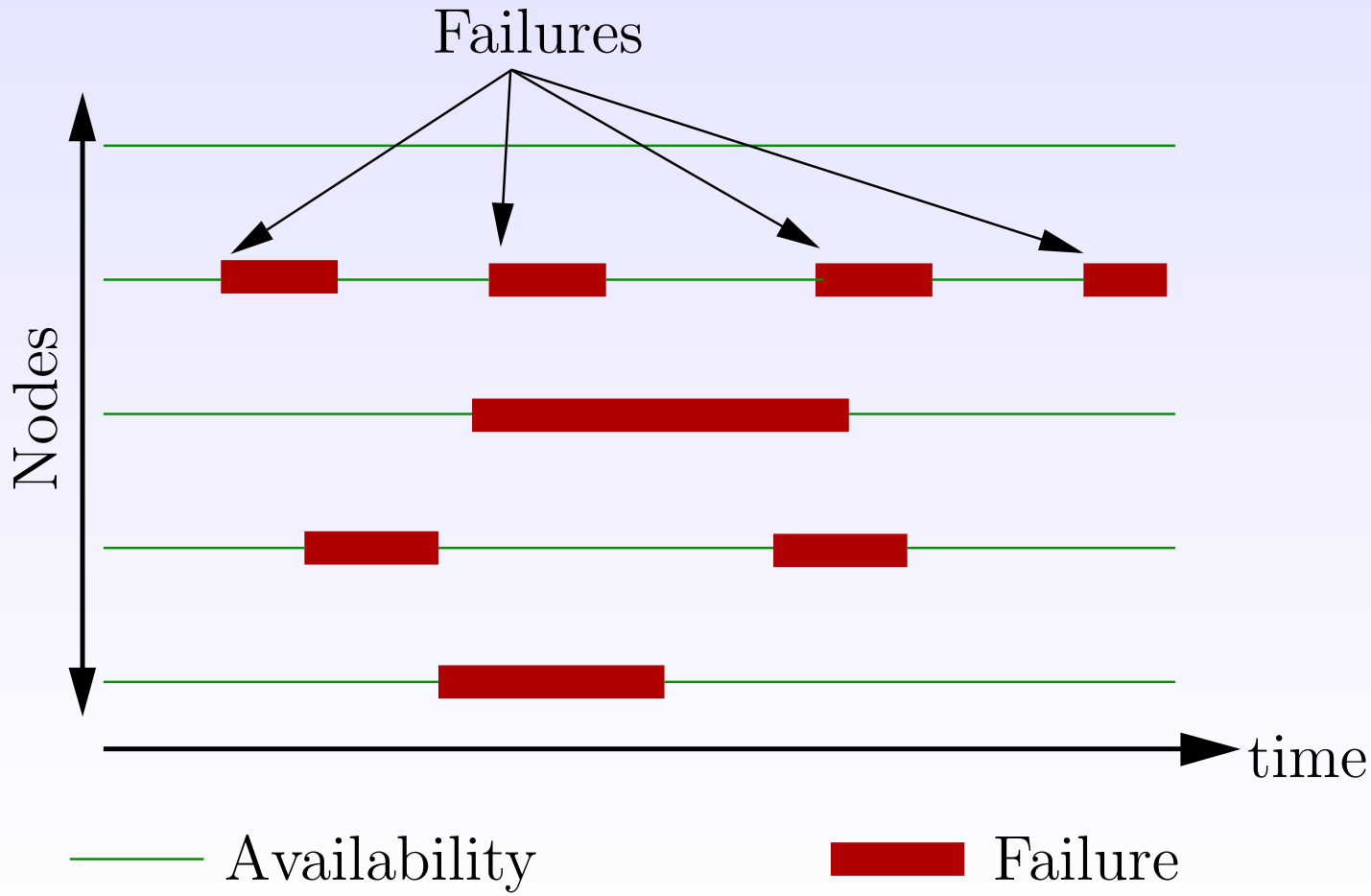
- Public repository of availability traces of parallel and distributed systems
- Records all unavailability and availability events
- Common SQL format for all traces

Definition

- Failure: unavailability event
- Availability event: return of the system to a correct state

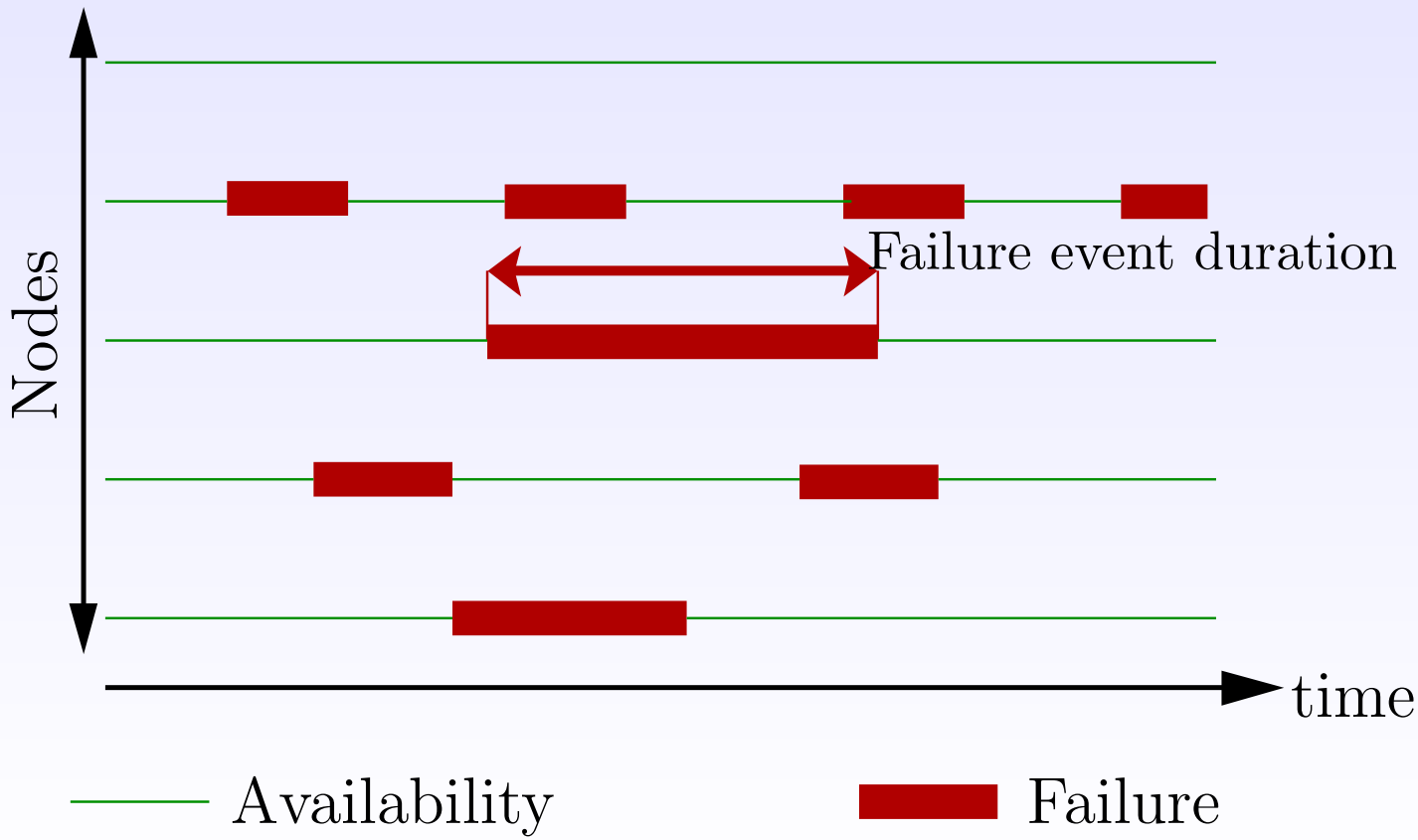
Definition

- Failure: unavailability event
- Availability event: return of the system to a correct state



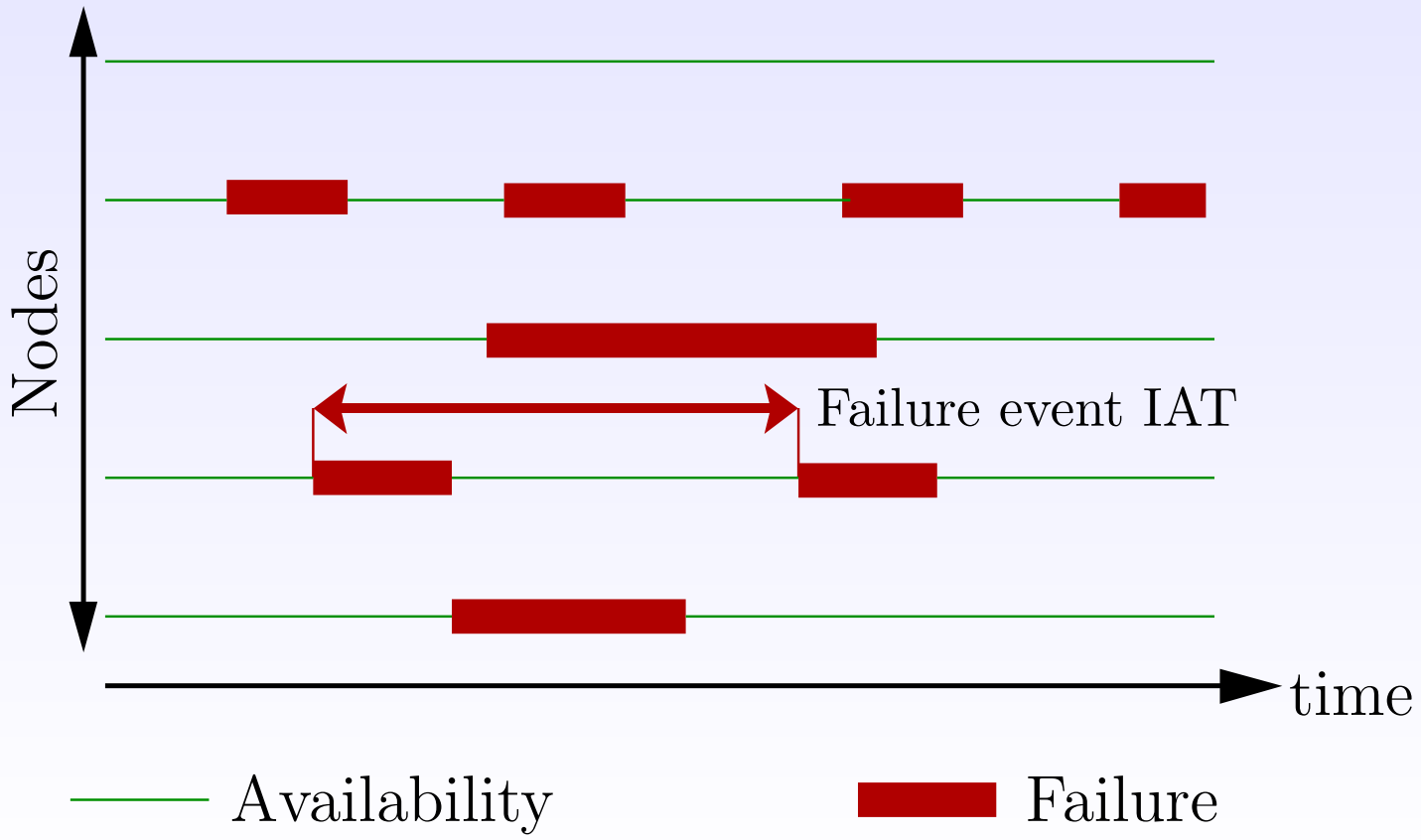
Definition

- Failure: unavailability event
- Availability event: return of the system to a correct state



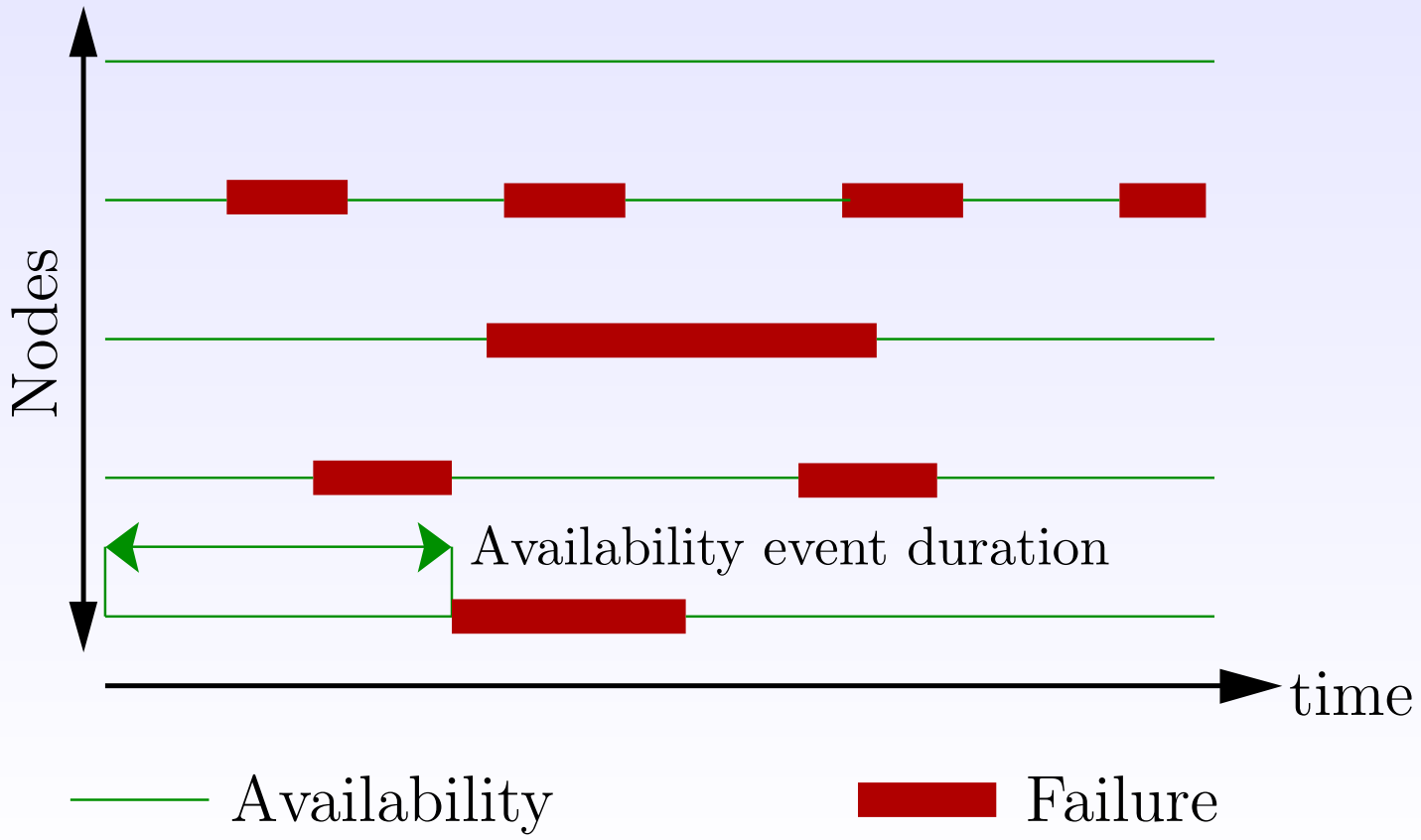
Definition

- Failure: unavailability event
- Availability event: return of the system to a correct state



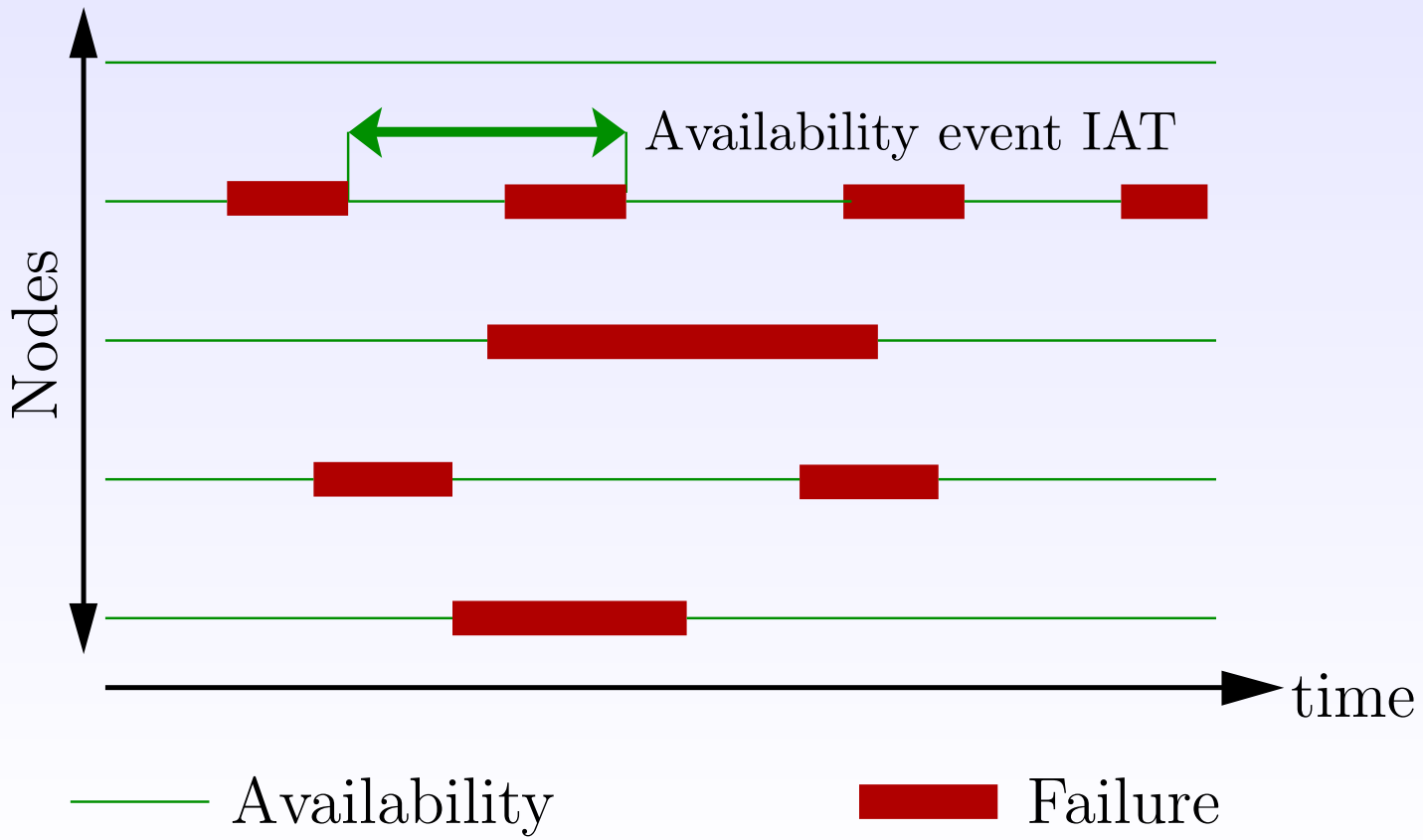
Definition

- Failure: unavailability event
- Availability event: return of the system to a correct state



Definition

- Failure: unavailability event
- Availability event: return of the system to a correct state



Analysis tools

- MySQL database (entire dumps are freely downloadable)
- Python scripts to select data and create graphs
- Matlab scripts to compute statistics and fitting
- Automated generation of complete reports (TeX and HTML files)
- `http://graal.ens-lyon.fr/~mgallet/Aleae/`

Datasets

System	Type	# Nodes	Period	Year	# of Events
Grid'5000	Grid	1,288	1.5 years	2005	588,463
Websites	Web servers	129	8 months	2001	95,557
LDNS	DNS servers	62,201	2 weeks	2004	384,991
LRI	Desktop Grid	237	10 days	2005	1,792
Deug	Desktop Grid	573	9 days	2005	33,060
SDSC	Desktop Grid	207	12 days	2003	6,882
UCB	Desktop Grid	80	11 days	1994	21,505
LANL	HPC Clusters	4,750	9 years	1996	43,325
Microsoft	Desktop	51,663	35 days	1999	1,019,765
PlanetLab	P2P	200-400	1.5 year	2004	49,164
Overnet	P2P	3,000	2 weeks	2003	68,892
Notre-Dame [1]	Desktop Grid	700	6 months	2007	300,241
Notre-Dame [2]	Desktop Grid	700	6 months	2007	268,202
Skype	P2P	4,000	1 month	2005	56,353
SETI	Desktop Grid	226,208	1.5 years	2007	202,546,160

[1] This is the host availability version which is according to the multi-state availability model of Brent Rood.

[2] This is the CPU availability version.

Space-Correlated Failures

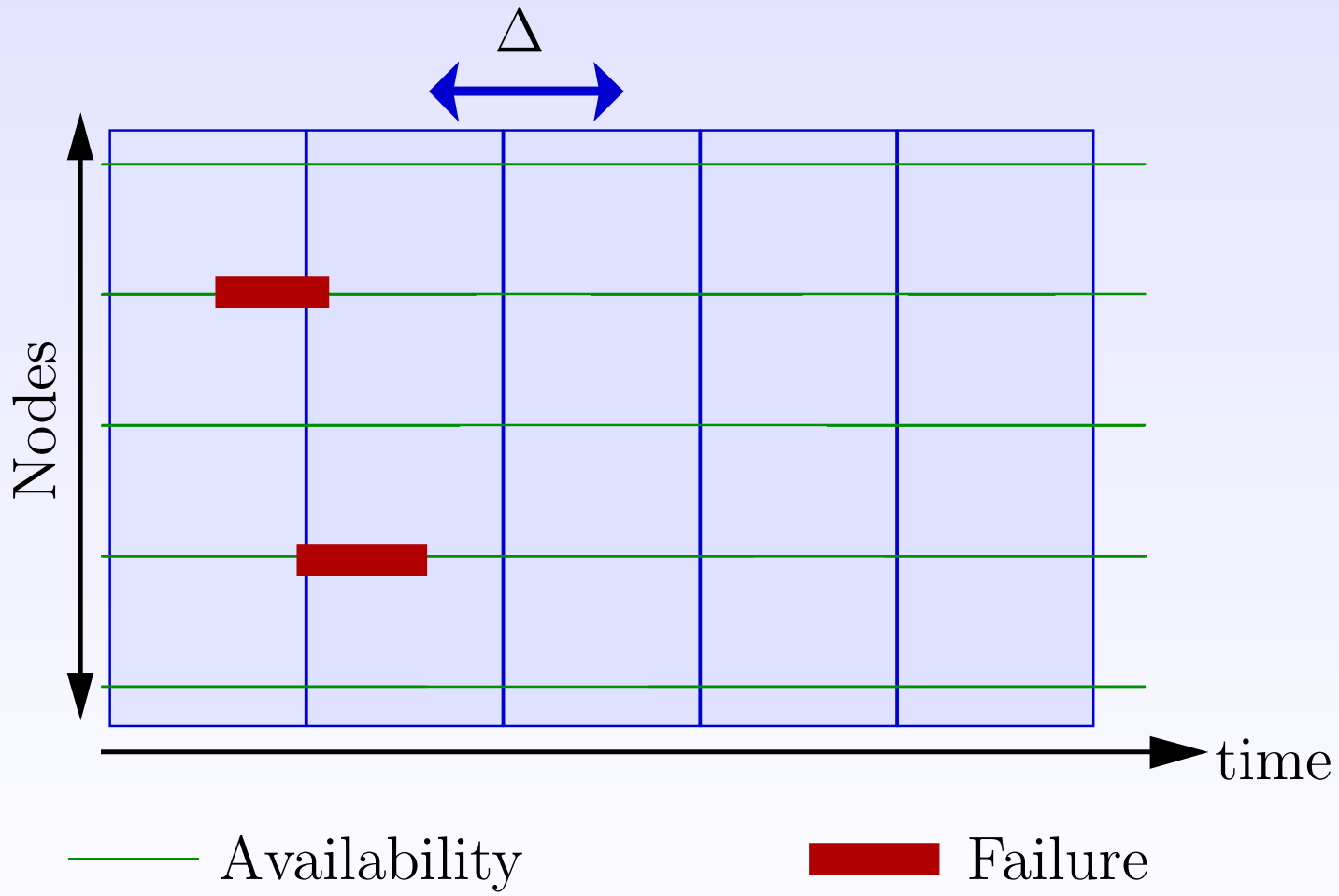
- Groups of failures occuring within a short time interval
- Lack of information in failure traces
- Numerical approach to identify groups of failures
- Three different approaches
- $TS(\cdot)$: time of an event

Space-Correlated Failures

- Groups of failures occuring within a short time interval
- Lack of information in failure traces
- Numerical approach to identify groups of failures
- Three different approaches
- $TS(\cdot)$: time of an event

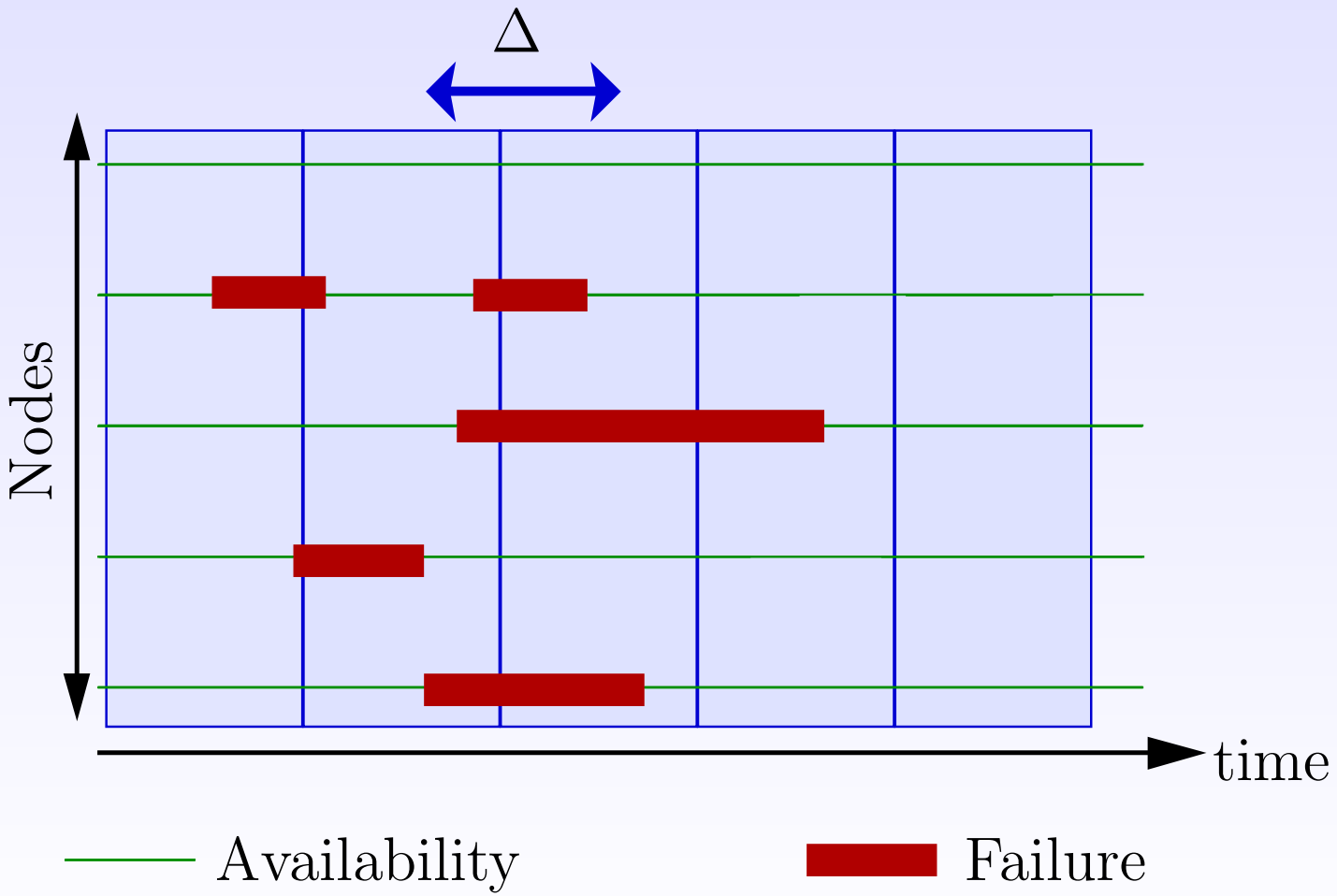
Time partitionning

- Partition time in windows of fixed size Δ



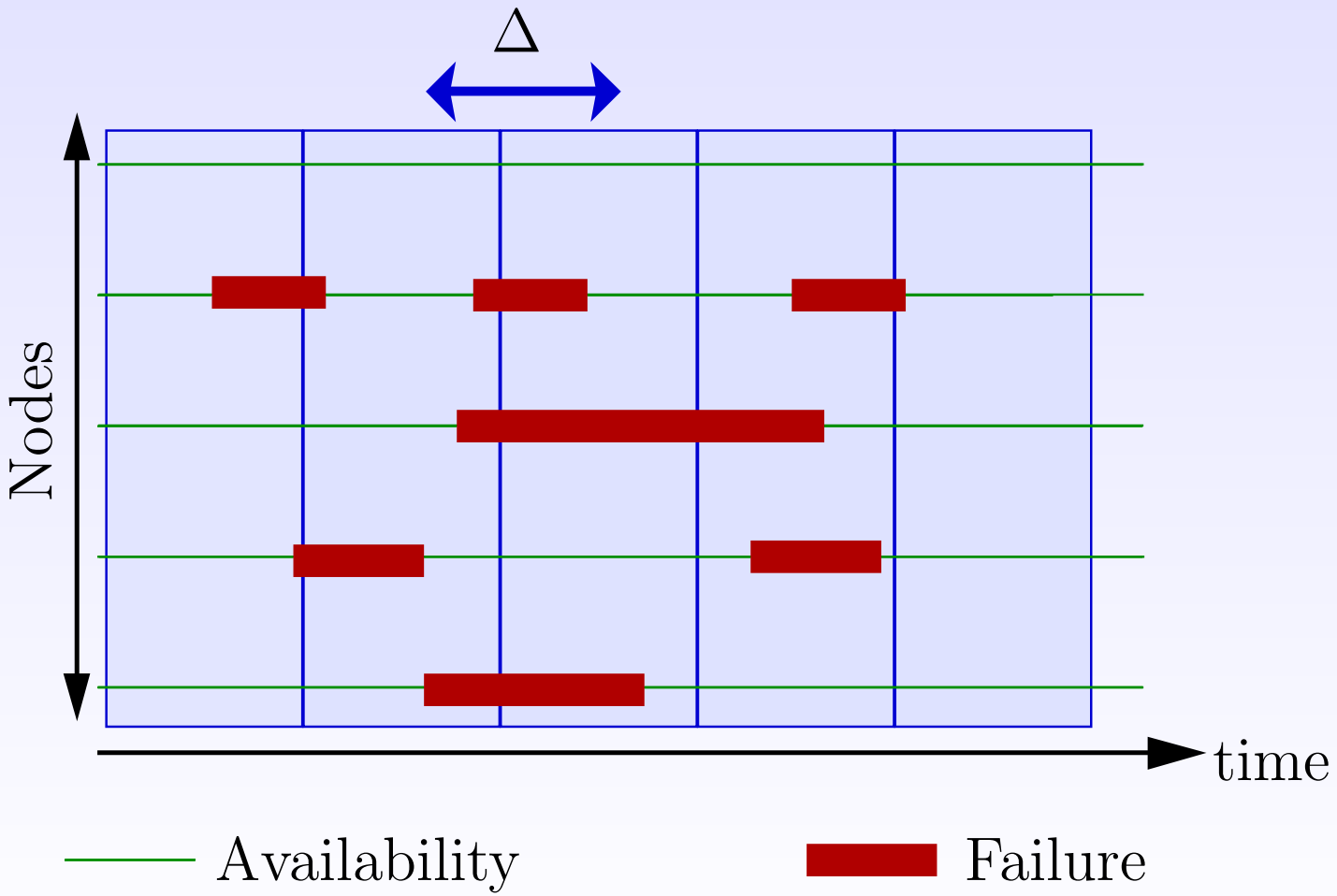
Time partitionning

- Partition time in windows of fixed size Δ



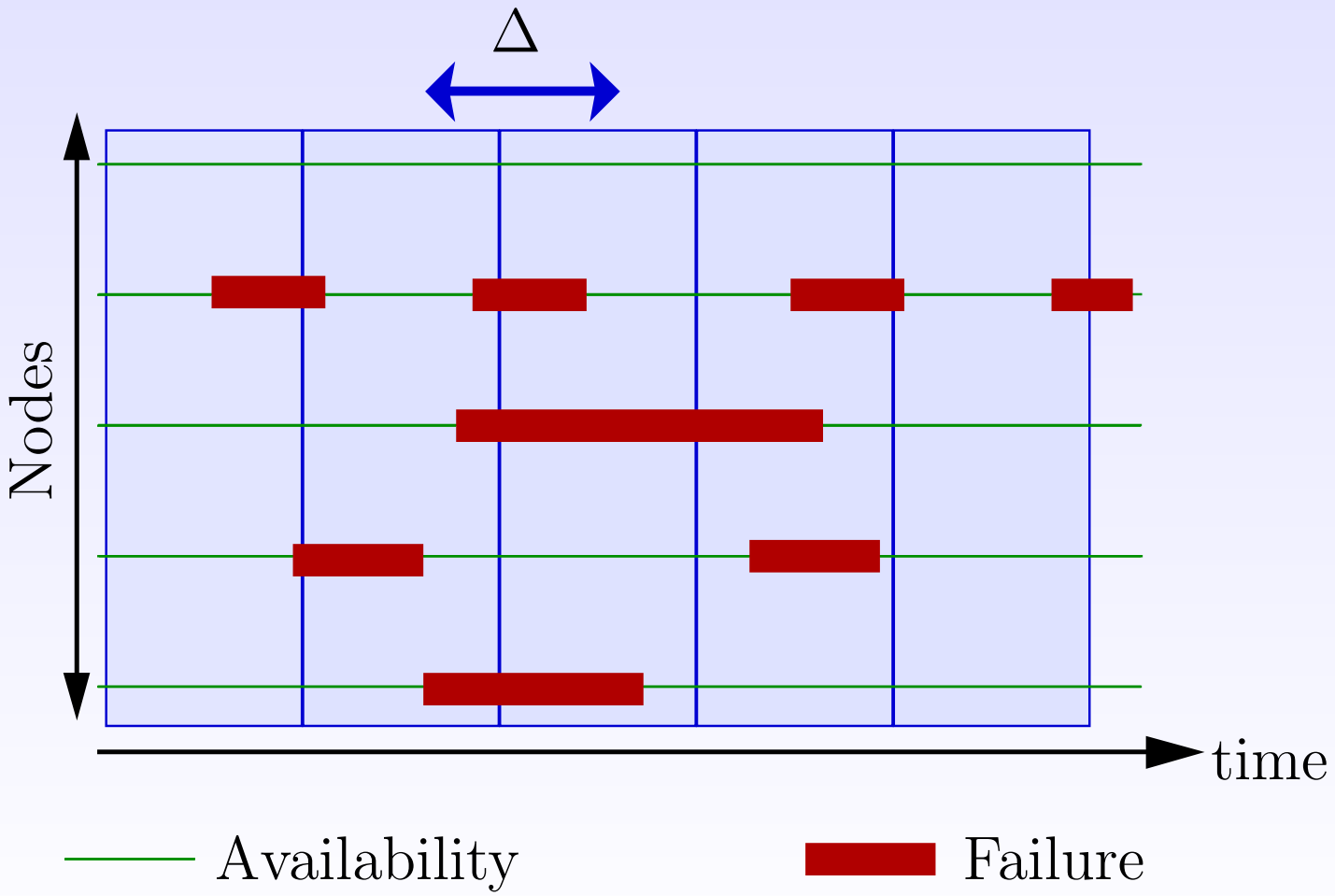
Time partitionning

- Partition time in windows of fixed size Δ



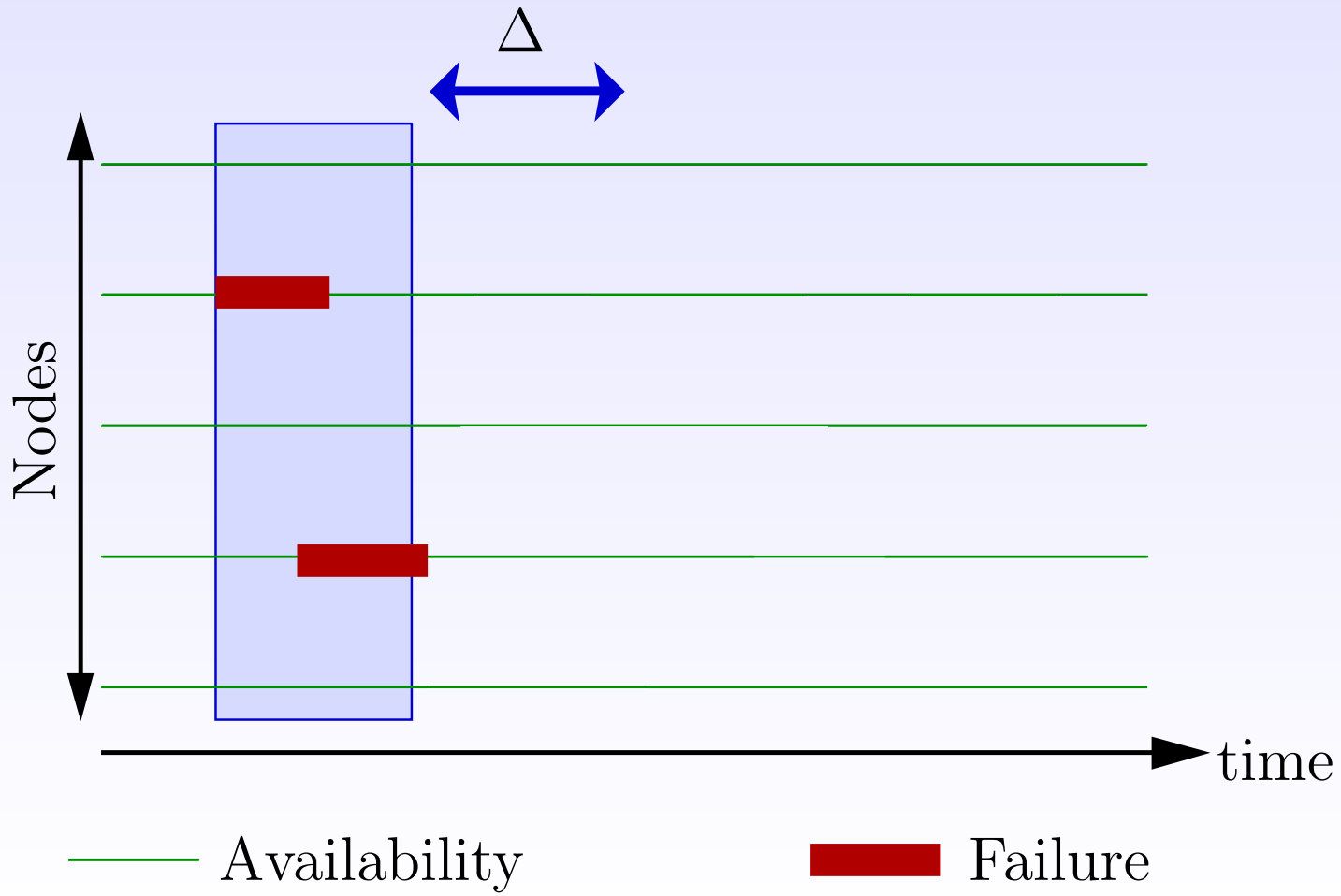
Time partitionning

- Partition time in windows of fixed size Δ



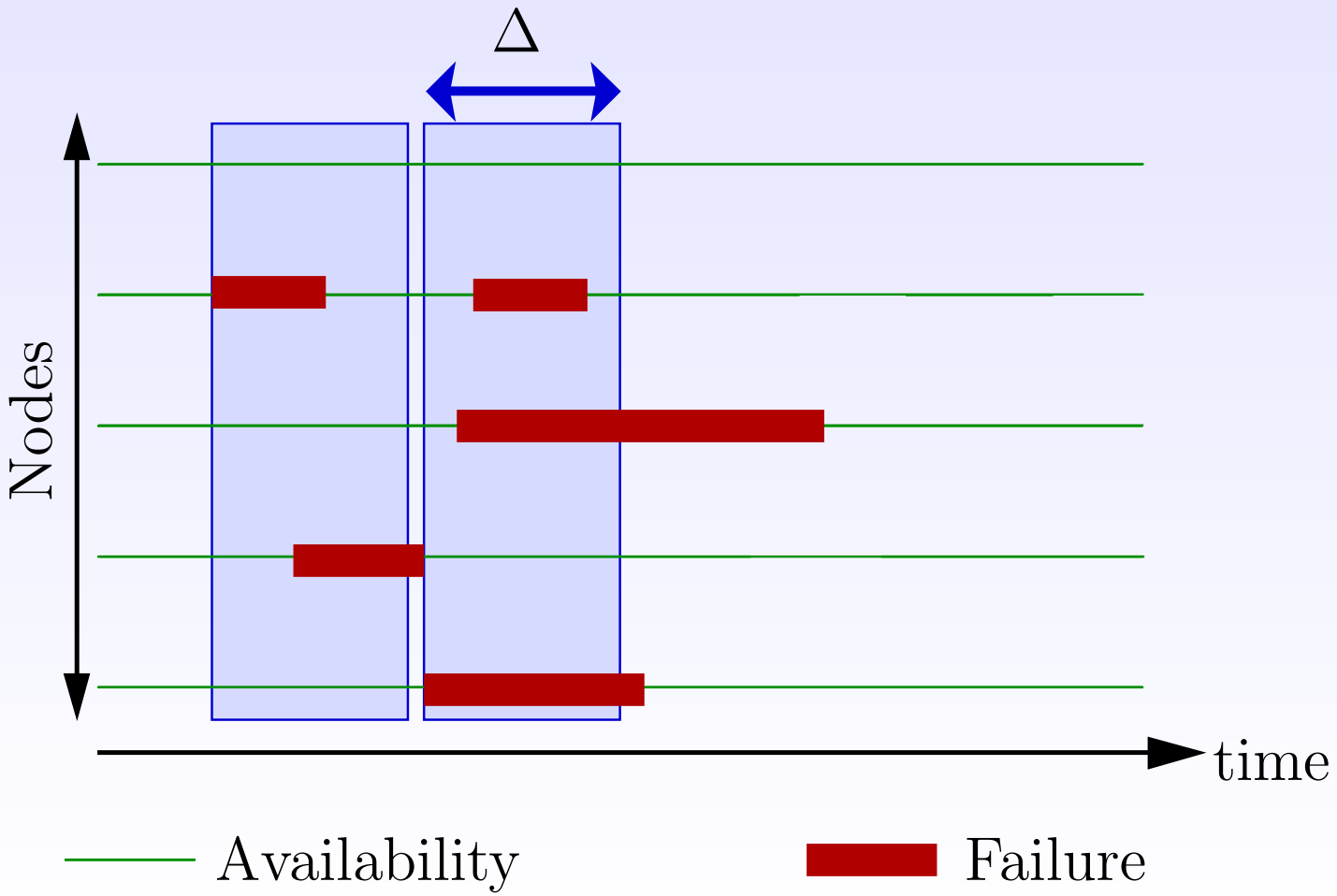
Moving windows

- The first isolated failure F_0 define a new failure group
- All failures within $TS(F_0)$ and $TS(F_0) + \Delta$ belong to this group



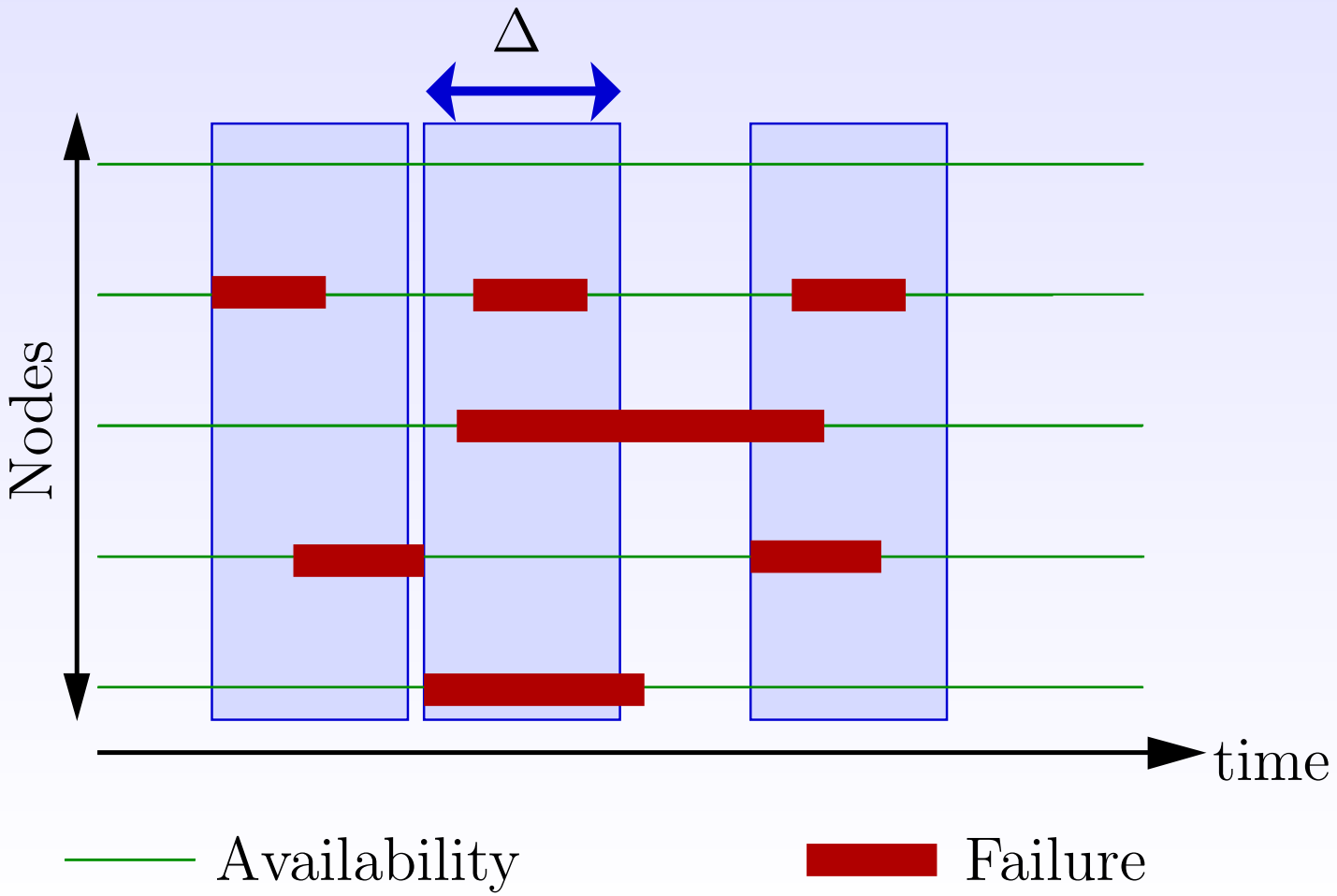
Moving windows

- The first isolated failure F_0 define a new failure group
- All failures within $TS(F_0)$ and $TS(F_0) + \Delta$ belong to this group



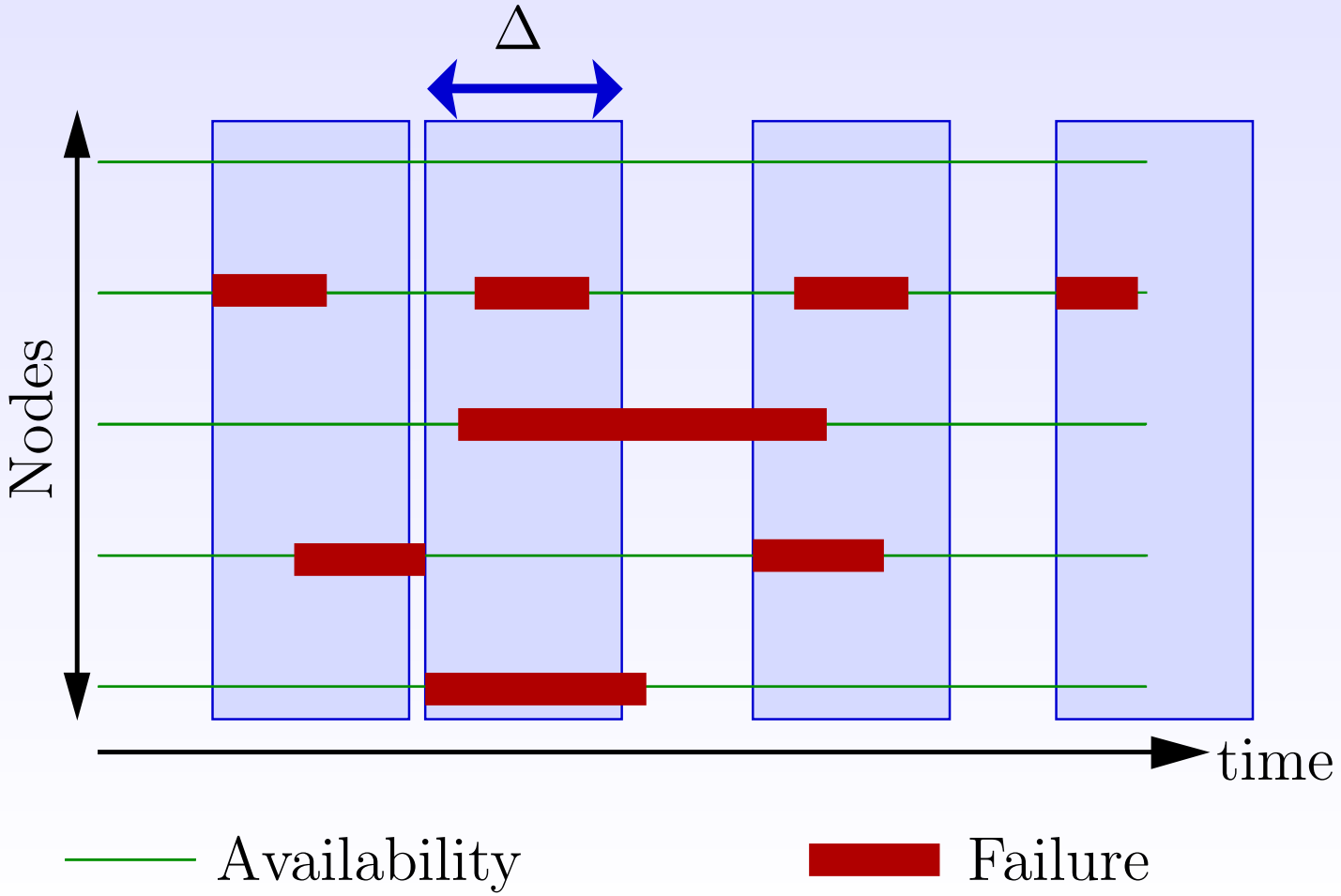
Moving windows

- The first isolated failure F_0 define a new failure group
- All failures within $TS(F_0)$ and $TS(F_0) + \Delta$ belong to this group



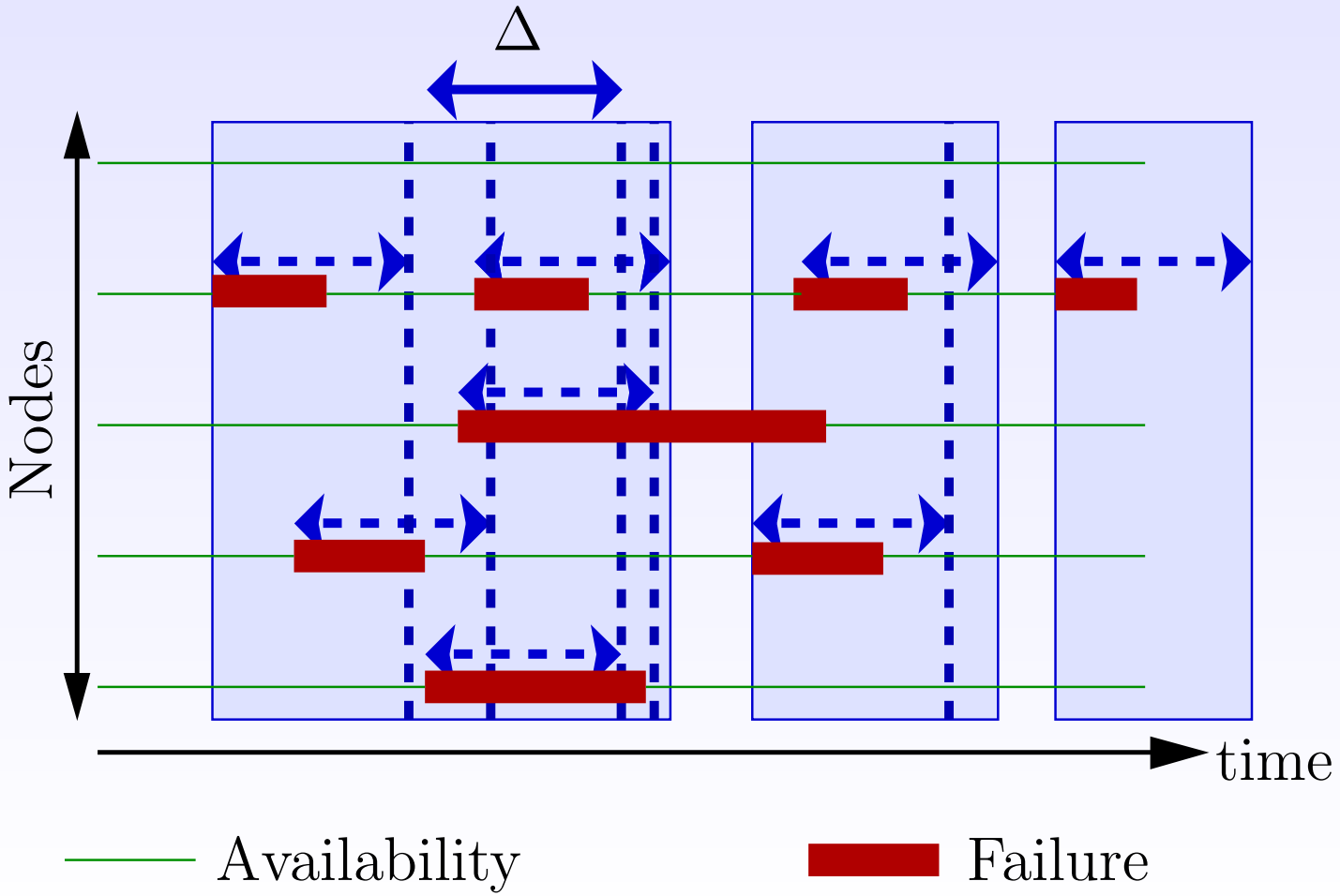
Moving windows

- The first isolated failure F_0 define a new failure group
- All failures within $TS(F_0)$ and $TS(F_0) + \Delta$ belong to this group



Extending windows

- The first isolated failure F define a new failure group
- Deadline is computed from the last failure of the group: $TS(F_i) \leq \max_{F_i} TS(F_i) + \Delta$



Model components

Inter-arrival Time : time between two successive groups

Size : number of failures in a group

Parallel Job Downtime : number of affected jobs times the duration

Single-node Job Downtime : sum of failure durations

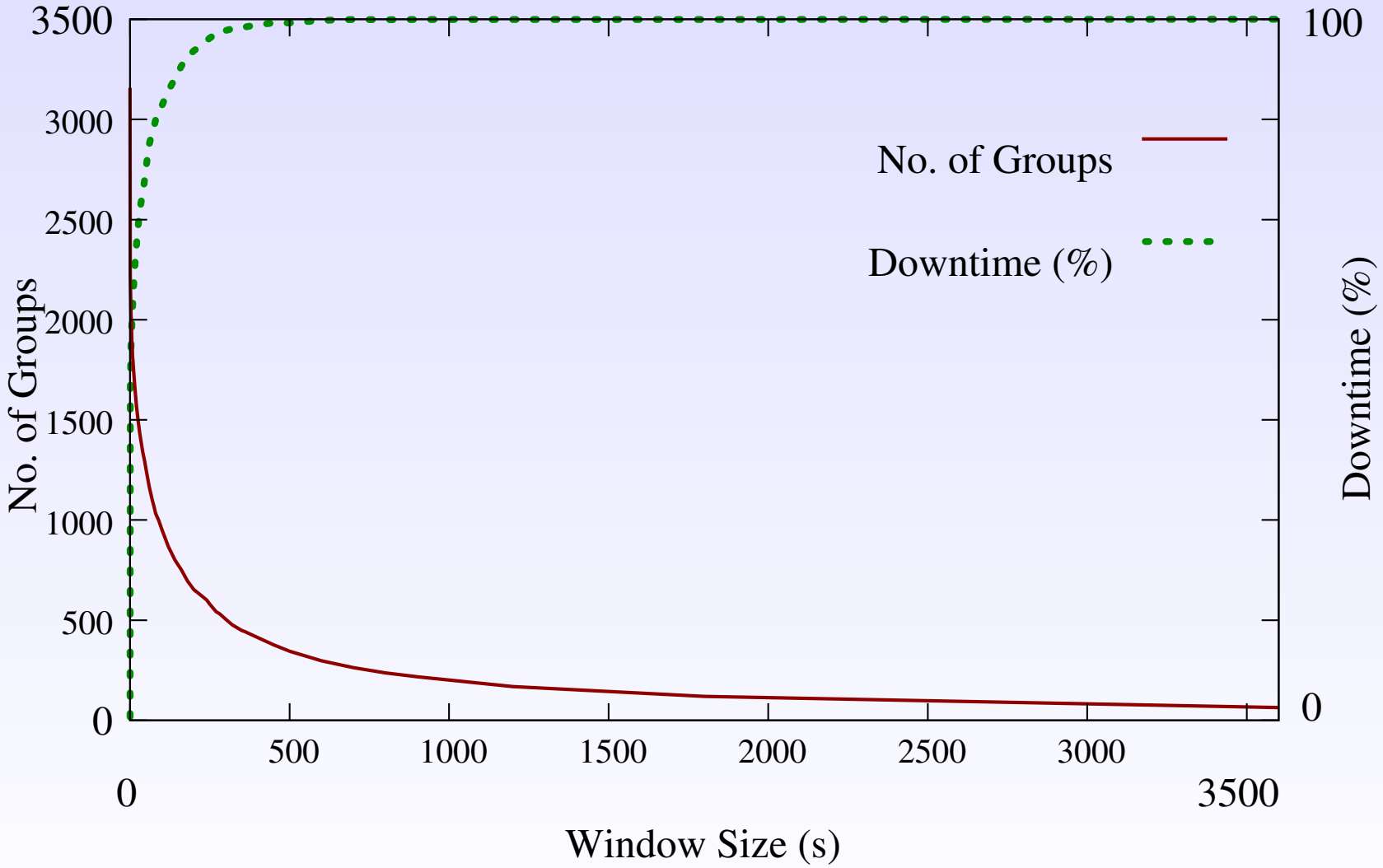
Method for Modeling

1. Analyze space-correlated failures for each trace
2. Characterize the properties of the empirical distributions
3. Find good fits for all distributions among several classical distributions
4. Assess the quality of these fits

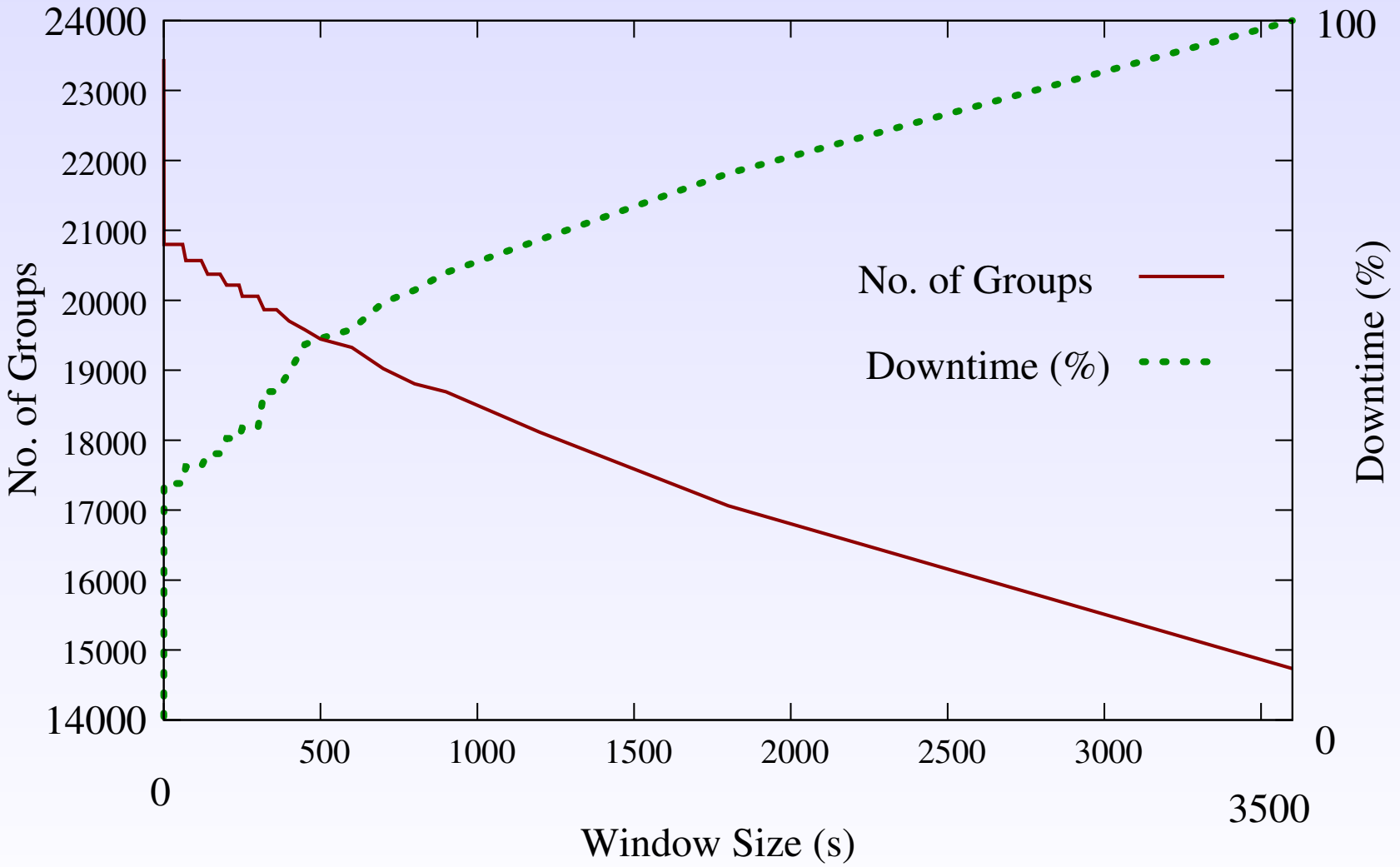
Failure Group Window Size

- Definition of failure groups relies on Δ
- Large values for Δ:
  - larger groups
  - increasing then decreasing number of groups of size at least 2
  - decreasing number of groups
  - reduced predictivity
- Small values for Δ:
  - smaller groups
  - few groups of size at least 2
  - convert to individual failures model

Failure Group Window Size: SDSC Platform



Failure Group Window Size: LANL Platform



Failure Group Window Size: Platform Selection

- Grid'5000
- Websites
- LDNS
- LRI
- Deug
- SDSC
- UCB

Analysis Results: Statistics I

Failure Group Inter-Arrival Time (hours)

	DEUG	GRID'5000	LDNS	LRI	SDSC	UCB	WEBSITES
Min	0.0419	0.069	0.0417	0.0278	0.0333	0.0222	0.0278
Q1	0.0504	0.103	0.0469	0.0941	0.0461	0.0242	0.169
Median	0.061	0.191	0.0608	0.183	0.0623	0.0272	0.829
Avg	1.220	0.529	0.178	0.865	0.468	0.515	11.850
Q3	0.112	0.444	0.0904	0.453	0.0900	0.0331	5.339
Max	84.025	58.644	6.495	77.462	64.094	65.005	1196.55
StdDev	7.644	1.275	0.714	4.818	4.039	3.946	40.102
COV	6.263	2.411	3.999	5.570	8.627	7.667	3.384
Sum	159.875	12704.4	15.162	237.837	279.982	213.053	566949
IQR	0.0615	0.341	0.0436	0.359	0.0439	0.00889	5.17
Skewness	9.810	12.177	8.219	14.697	13.683	11.955	9.020
Kurtosis	105.294	329.899	72.357	232.515	208.224	178.252	135.887

Analysis Results: Statistics II

Failure Group Size

	DEUG	GRID'5000	LDNS	LRI	SDSC	UCB
Min	2	2	2	2	2	2
Q1	5	2	10	2	2	3
Median	9	4	13	3	3	4
Avg	10.962	17.095	13.440	5.739	5.194	4.472
Q3	14	16	16	5	4	6
Max	157	600	82	41	165	14
StdDev	9.612	33.546	5.918	8.129	11.882	1.873
COV	0.877	1.962	0.441	1.417	2.288	0.419
Sum	16388	283275	161372	505	2836	10385
Kurtosis	74.892	36.644	18.768	12.128	96.757	4.451
Skewness	6.302	4.664	2.619	3.142	8.794	0.990
IQR	9	14	6	3	2	3

Analysis Results: Statistics III

Single-node Job Downtime

	DEUG	GRID'5000	LDNS	LRI	SDSC	UCB	WEBSITES
Min	191	12	1038.48	104.014	282.032	36	1168
Q1	4262.37	1106	732346	5181.15	8884.62	1215	1330
Median	20238.8	4614	1.51e6	72279.1	19423.9	2658	3670
Avg	117911	3.33e6	2.48e6	246482	67183.3	4071.25	21225.2
Q3	149560	128430	2.95e6	167646	33806.1	5220	15674
Max	6.36e6	2.03e9	5.67e7	3.47e6	4.75e6	65584	2.38e7
StdDev	346344	3.13e7	3.20e6	622008	296050	4718.94	332737
COV	2.937	9.387	1.286	2.523	4.406	1.160	15.676
Sum	1.76e8	5.52e10	2.98e10	2.17e7	3.67e7	9.45e6	1.95e8
Kurtosis	172.637	1289.01	35.759	19.658	134.924	34.752	3914.84
Skewness	11.774	27.851	4.313	4.064	10.253	4.124	60.226
IQR	145298	127324	2.22e6	162464	24921.5	4005	14344

Analysis Results: Statistics IV

Parallel Job Downtime

	Deug	Grid'5000	LDNS	LRI	SDSC	UCB	Websites
Min	160	11	1042	103	182.387	22	1160
Q1	1628	837	193345	4318.64	5628.18	662	1189
Median	5089	3125	335095	52752.1	11453.1	1179	2400
Avg	29979.3	440266	416568	162946	30139.7	1500.92	10363.5
Q3	20956.3	55739	542486	145670	18684.4	1976	8400
Max	3.27e6	3.89e8	3.91e6	2.04e6	3.51e6	12365	1.19e7
StdDev	178430	4.91e6	327308	356392	165727	1226.51	137441
COV	5.952	11.157	0.786	2.187	5.499	0.817	13.262
Sum	4.48e7	7.30e9	5.00e9	1.43e7	1.65e7	3.49e6	9.50e7
Kurtosis	214.493	3010.62	10.982	17.964	360.761	12.345	6246.92
Skewness	14.092	46.638	2.103	3.765	17.758	2.272	74.742
IQR	19328.3	54902	349140	141351	13056.2	1314	7211

Analysis Results: Fitting I

Failure Group Inter-Arrival Time (hours)

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	0.23	0.12	0.18	0.86	1.22	0.47	0.21
Weibull	0.44 0.79	0.16 1.21	0.12 0.74	0.46 0.63	0.23 0.47	0.13 0.27	0.07 0.48
Pareto	0.42 0.29	0.01 0.12	0.36 0.08	0.62 0.22	0.84 0.09	0.40 0.07	0.21 0.03
Logn	-1.39 1.03	-2.17 0.76	-2.27 0.81	-1.46 1.28	-2.28 1.32	-2.63 0.86	-3.41 0.98
Gamma	0.79 0.67	1.83 0.08	0.71 0.22	0.48 1.79	0.28 4.33	0.36 1.31	0.26 2.00

Analysis Results: Fitting II

Failure Group Size

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Exp	17.09	2.55	13.44	5.74	10.96	5.19	4.47
Weibull	12.82 0.71	2.87 1.60	15.12 2.29	5.76 1.01	12.12 1.39	4.94 0.93	5.05 2.52
Pareto	0.68 6.75	-0.06 2.68	-0.18 15.09	0.22 4.43	-0.03 11.26	0.22 3.70	-0.41 5.76
Logn	1.88 1.25	0.84 0.35	2.52 0.41	1.32 0.77	2.15 0.70	1.19 0.70	1.41 0.42
Gamma	0.64 26.78	5.33 0.48	6.23 2.16	1.30 4.40	2.22 4.94	1.23 4.24	6.03 0.74

Analysis Results: Fitting III

Single-node Job Downtime

	Grid'5000	Websites	LDNS	LRI
Exp	3.33e6	21225.18	2.48e6	2.46e5
Weibull	75972.13 0.28	10658.82 0.63	2.430e6 0.96	1.051e5 0.48
Pareto	3.10 2686.08	0.73 5493.50	0.16 2.071e6	1.71 24187.13
Logn	9.51 3.21	8.57 1.36	14.16 1.15	10.41 2.45
Gamma	0.14 2.362e6	0.46 46006.96	1.01 2.452e6	0.34 7.317e5

	Deug	SDSC	UCB
Exp	1.18e5	67183.25	4071.25
Weibull	61989.86 0.54	35581.34 0.63	4131.60 1.03
Pareto	1.53 15901.44	0.54 20627.60	0.09 3711.35
Logn	10.03 2.02	9.80 1.30	7.82 1.03
Gamma	0.40 2.950e5	0.49 1.384e5	1.16 3509.88

Analysis Results: Fitting IV

Parallel Job Downtime

	Grid'5000	Websites	LDNS	LRI
Exp	4.40e5	1.04e5	4.17e5	1.63e5
Weibull	3.10e5 0.33	6605.36 0.70	4.576e5 1.37	8.01e5 0.50
Pareto	2.54 2215.71	0.47 4258.00	-0.11 4.576e5	1.61 2.07e5
Logn	8.89 2.71	8.20 1.13	12.64 0.84	10.16 2.40
Gamma	0.18 2.42e6	0.59 1.75e5	1.82 2.29e5	0.36 4.48e5

	Deug	SDSC	UCB
Exp	3.00e5	3.01e5	1500.92
Weibull	1.32e5 0.57	1.90e5 0.69	1646.49 1.35
Pareto	0.91 5832.36	0.41 1.26e5	-0.10 1645.39
Logn	8.67 1.62	9.25 1.16	7.01 0.81
Gamma	0.40 7.50e5	0.59 5.15e5	1.82 825.92

Analysis Results: GoF I

Failure Group Inter-Arrival Time, both AD and KS tests. Values in red denote p-values above the test threshold for the significance level 0.05.

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Anderson-Darling test							
Exp	0.185	0.329	0.0175	0.0361	1.50e-5	1.25e-6	5.78e-7
Weibull	0.221	0.491	0.034	0.287	0.0242	0.0074	0.0047
Pareto	9.39e-5	1.09e-5	9.85e-7	6.56e-4	3.14e-6	2.57e-6	6.58e-8
Logn	0.446	0.607	0.183	0.569	0.0801	0.156	0.0153
Gamma	0.2	0.55	0.026	0.147	0.0072	0.0018	0.0012
Kolmogorov-Smirnov test							
Exp	0.0911	0.185	8.27e-4	0.0035	3.10e-10	3.21e-11	1.61e-17
Weibull	0.075	0.409	5.27e-4	0.153	4.550-4	2.33e-4	4.38e-6
Pareto	3.04e-6	1.96e-7	3.16e-8	7.98e-5	1.01e-8	1.36e-6	5.21e-11
Logn	0.333	0.466	0.0496	0.372	0.0148	0.104	6.34e-4
Gamma	0.0995	0.437	0.0021	0.0579	3.49e-4	3.04e-5	1.67e-7

Analysis Results: GoF II

Failure Group Size, both AD and KS tests. Values in red denote p-values above the test threshold for the significance level 0.05.

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Anderson-Darling test							
Exp	0.28	0.439	0.14	0.508	0.449	0.398	0.333
Weibull	0.502	0.498	0.597	0.516	0.639	0.408	0.734
Pareto	0.122	0.244	4.86e-5	0.079	4.83e-4	0.115	0.0049
Logn	0.575	0.684	0.711	0.647	0.708	0.622	0.765
Gamma	0.393	0.595	0.699	0.469	0.687	0.392	0.772
Kolmogorov-Smirnov test							
Exp	0.0232	1.49e-7	0.0011	0.0072	0.119	0.0026	3.41e-4
Weibull	0.0485	1.40e-6	0.257	0.008	0.418	8.32e-4	0.206
Pareto	1.19e-5	6.24e-14	9.14e-11	2.18e-11	2.09e-6	1.88e-8	4.74e-14
Logn	0.144	2.69e-4	0.37	0.0651	0.421	0.0267	0.174
Gamma	0.0578	2.16e-4	0.376	0.024	0.44	0.0073	0.191

Analysis Results: GoF III

Single-node Job Downtime, both AD and KS tests. Values in bold denote p-values above the test threshold for the significance level 0.05.

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Anderson-Darling test							
Exp	1.18e-6	0.0364	0.563	0.0643	0.047	0.0151	0.565
Weibull	0.235	0.278	0.556	0.606	0.513	0.23	0.578
Pareto	0.0118	5.99e-4	3.84e-5	0.0593	0.0061	2.74e-4	1.34e-5
Logn	0.368	0.356	0.61	0.609	0.548	0.534	0.618
Gamma	0.0528	0.143	0.558	0.493	0.428	0.0962	0.571
Kolmogorov-Smirnov test							
Exp	5.41e-11	0.0083	0.004	0.0076	0.0085	0.0021	0.465
Weibull	0.107	0.0733	0.303	0.292	0.348	0.111	0.477
Pareto	0.0019	1.46e-5	0.0017	0.0035	0.0011	6.10e-5	5.27e-7
Logn	0.223	0.223	0.263	0.253	0.347	0.383	0.481
Gamma	0.023	0.0709	0.251	0.236	0.301	0.0291	0.464

Analysis Results: GoF IV

Parallel Job Downtime, both AD and KS tests. Values in bold denote p-values above the test threshold for the significance level 0.05.

	Grid'5000	Websites	LDNS	LRI	Deug	SDSC	UCB
Anderson-Darling test							
Exp	4.85e-5	0.0903	0.396	0.135	0.0189	0.0632	0.405
Weibull	0.26	0.222	0.592	0.627	0.455	0.243	0.59
Pareto	0.0086	3.17e-4	1.23e-6	0.0599	0.0017	9.12e-5	5.02e-6
Logn	0.393	0.342	0.602	0.583	0.576	0.535	0.626
Gamma	0.0775	0.158	0.606	0.547	0.23	0.139	0.602
Kolmogorov-Smirnov test							
Exp	1.67e-7	0.0299	0.233	0.0209	0.004	0.0149	0.276
Weibull	0.138	0.0304	0.48	0.325	0.37	0.136	0.485
Pareto	0.0014	5.06e-6	1.92e-7	0.0033	1.78e-4	1.64e-5	1.13e-7
Logn	0.24	0.183	0.471	0.235	0.414	0.373	0.494
Gamma	0.0372	0.0569	0.492	0.293	0.177	0.0527	0.473

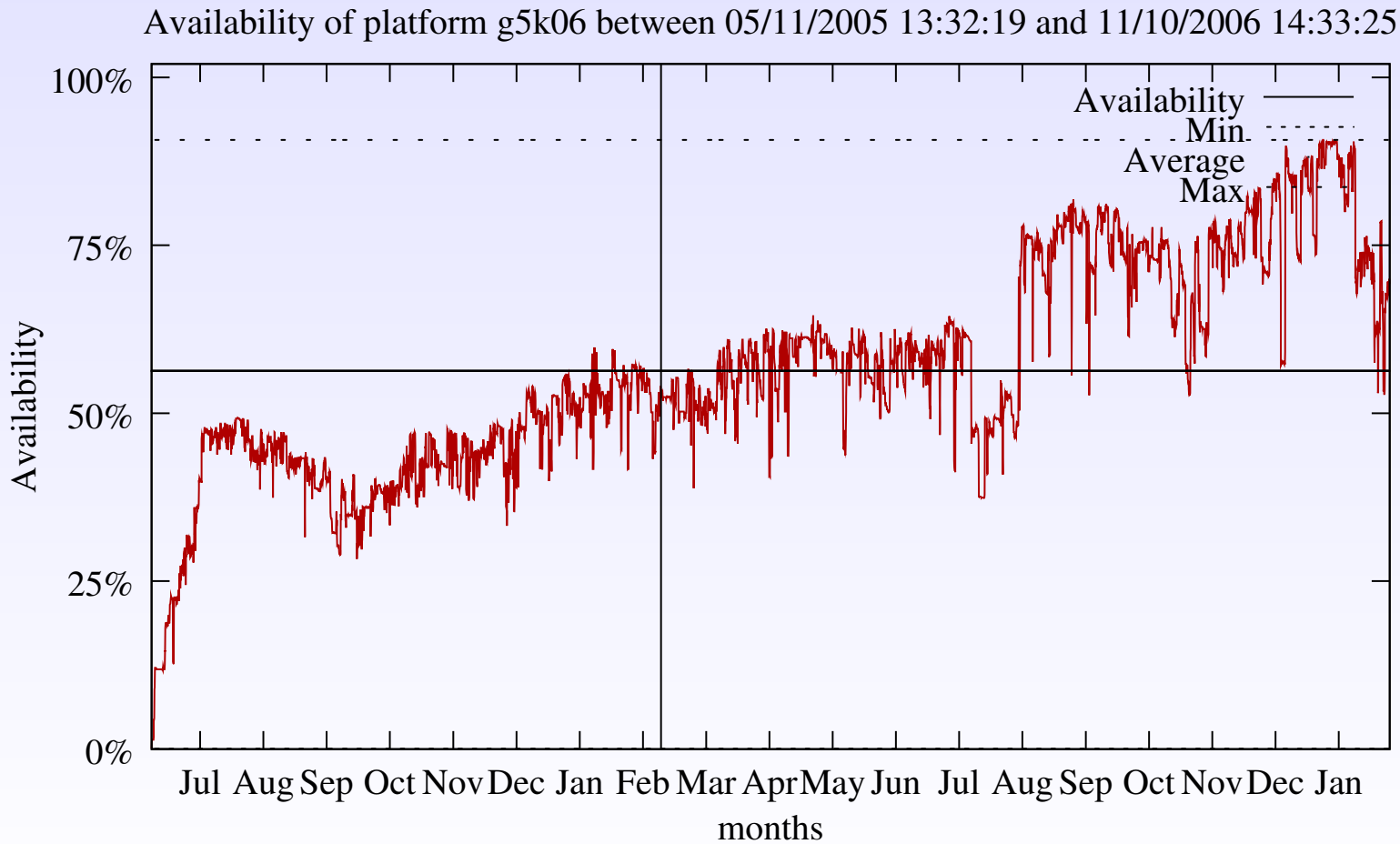
Results summary

	Group size	Group IAT	$D_{\max}$	D_{Σ}
Grid'5000	Logn (1.88,1.25)	Logn (-1.39,1.03)	Logn (9.51,3.21)	Logn (8.89,2.71)
Websites	Gamma (0.84,0.35)	Logn (-2.17,0.76)	Logn (8.57,1.36)	Logn (8.20,1.13)
LDNS	Logn (2.52,0.41)	Logn (-2.57,0.81)	Logn (14.16,1.15)	Gamma (1.82,2.292e5)
LRI	Logn (1.32,0.77)	Logn (-1.46,1.28)	Weibull (1.051e5,0.48)	Weibull (80091.30,0.50)
Deug	Logn (2.15,0.70)	Logn (-2.28,1.35)	Logn (10.03,2.02)	Logn (8.67,1.62)
SDSC	Logn (1.10,0.70)	Logn (-2.63,0.86)	Logn (9.80,1.30)	Logn (9.25,1.16)
UCB	Gamma (6.03,0.74)	Logn (-3.41,0.98)	Logn (7.82,1.03)	Logn (7.01,0.81)

- Log-Normal distribution provides good fits for almost all components and platforms
- Most node-level failures can be modeled by groups of failures
- Few components are required to describe failures in large platforms

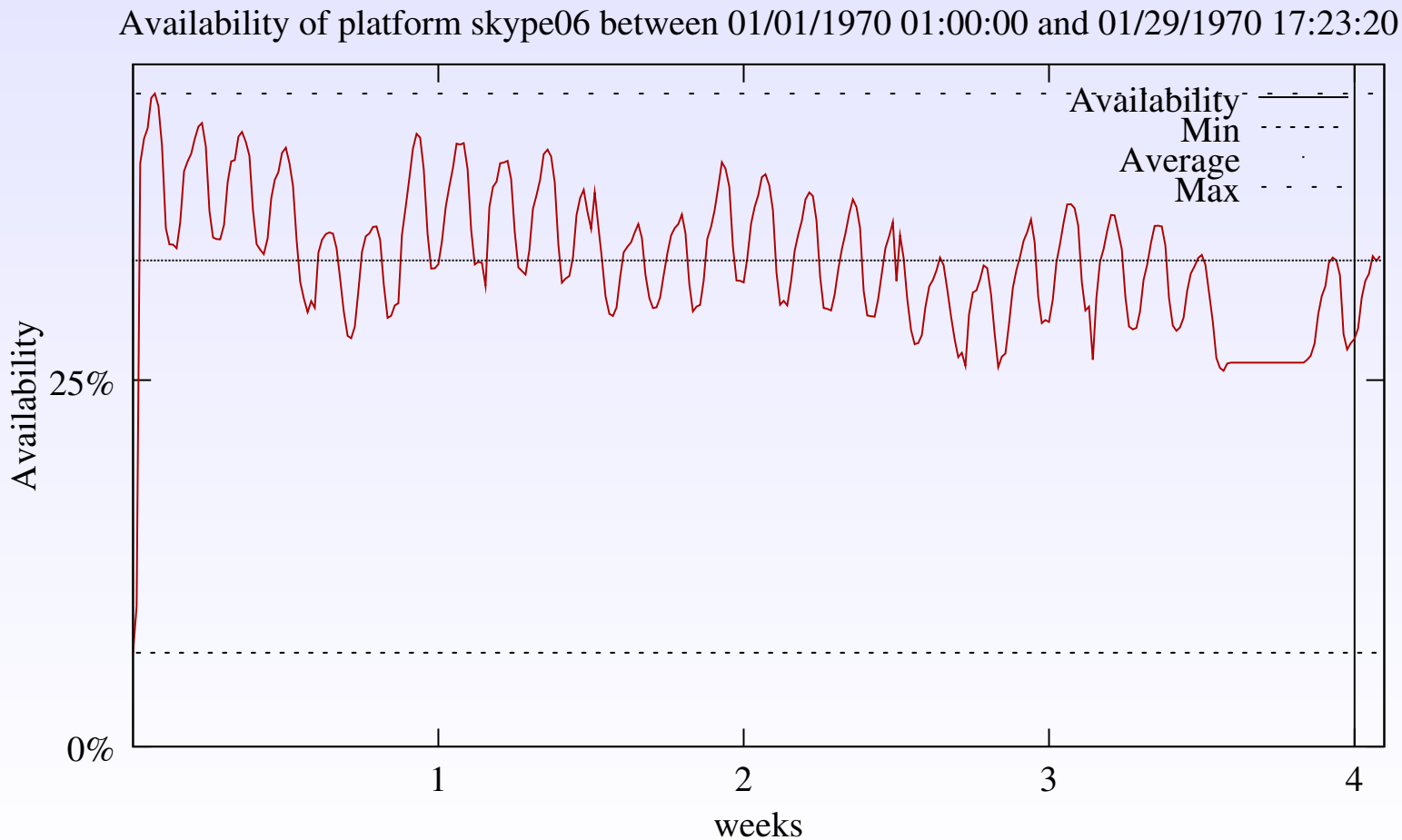
Future Work (I)

- New ways to model the evolution of platform availability
- Needs some new automated tools



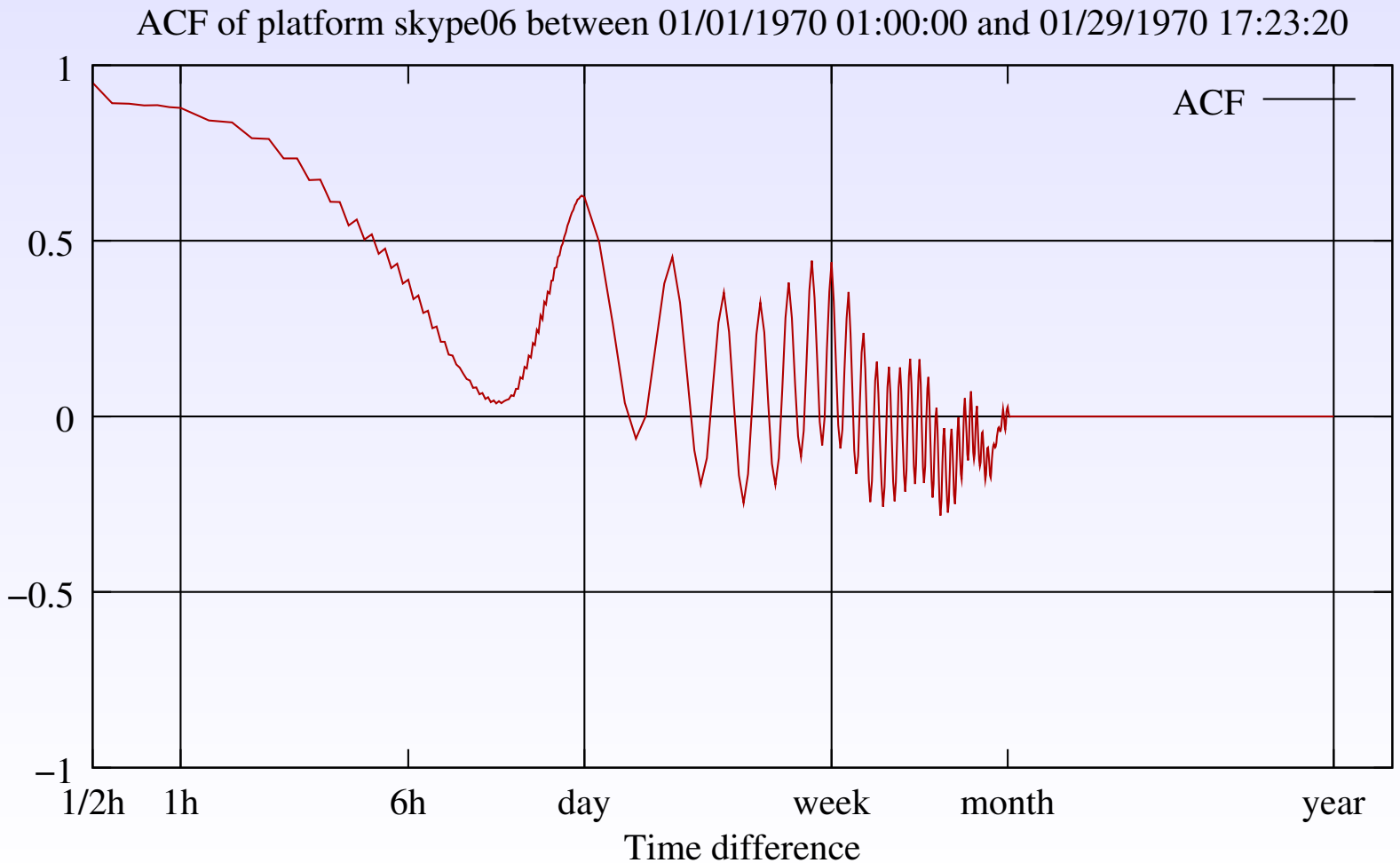
Future Work (II)

- Use of time patterns and autocorrelation function
- *Time-correlated* failures
- Separately analyze high-availability and low-availability periods



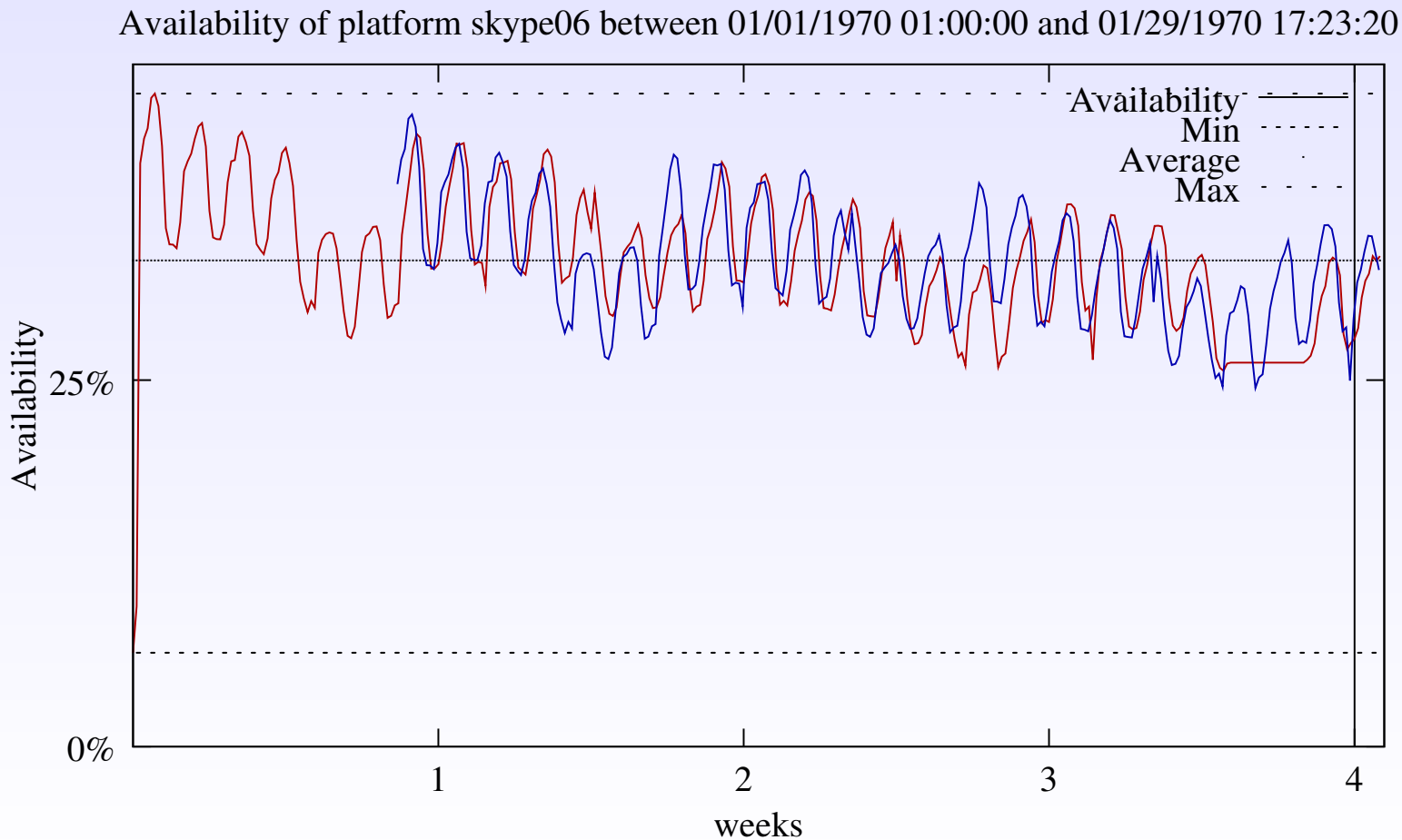
Future Work (II)

- Use of time patterns and autocorrelation function
- *Time-correlated* failures
- Separately analyze high-availability and low-availability periods



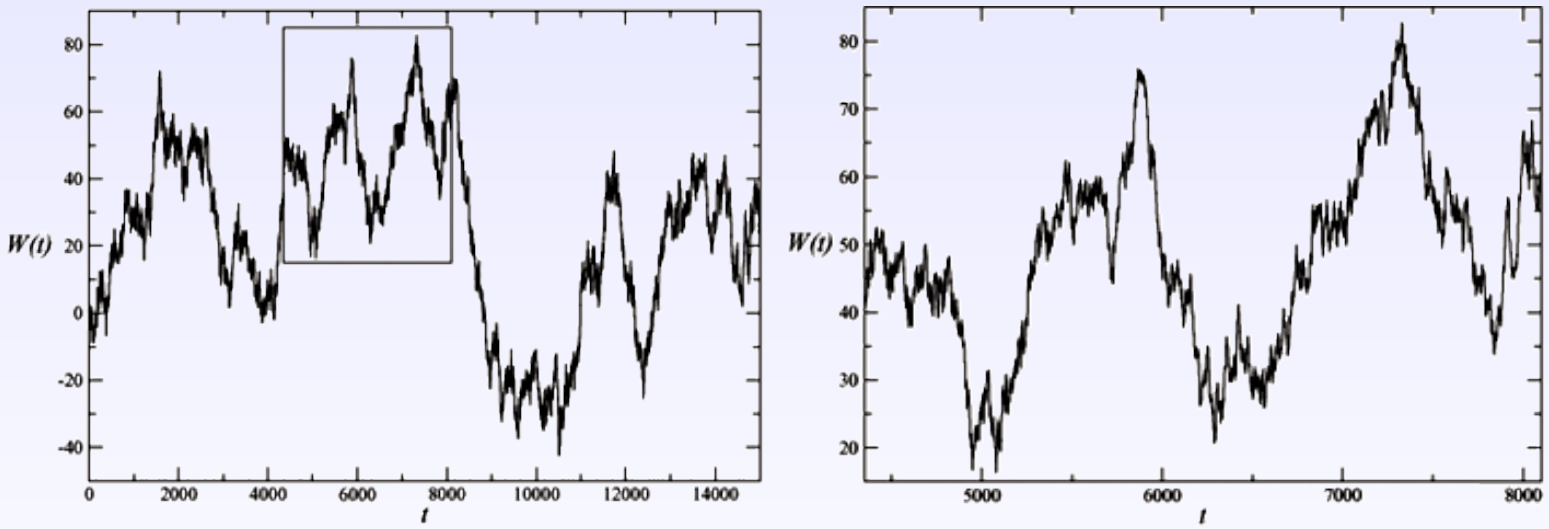
Future Work (II)

- Use of time patterns and autocorrelation function
- *Time-correlated* failures
- Separately analyze high-availability and low-availability periods



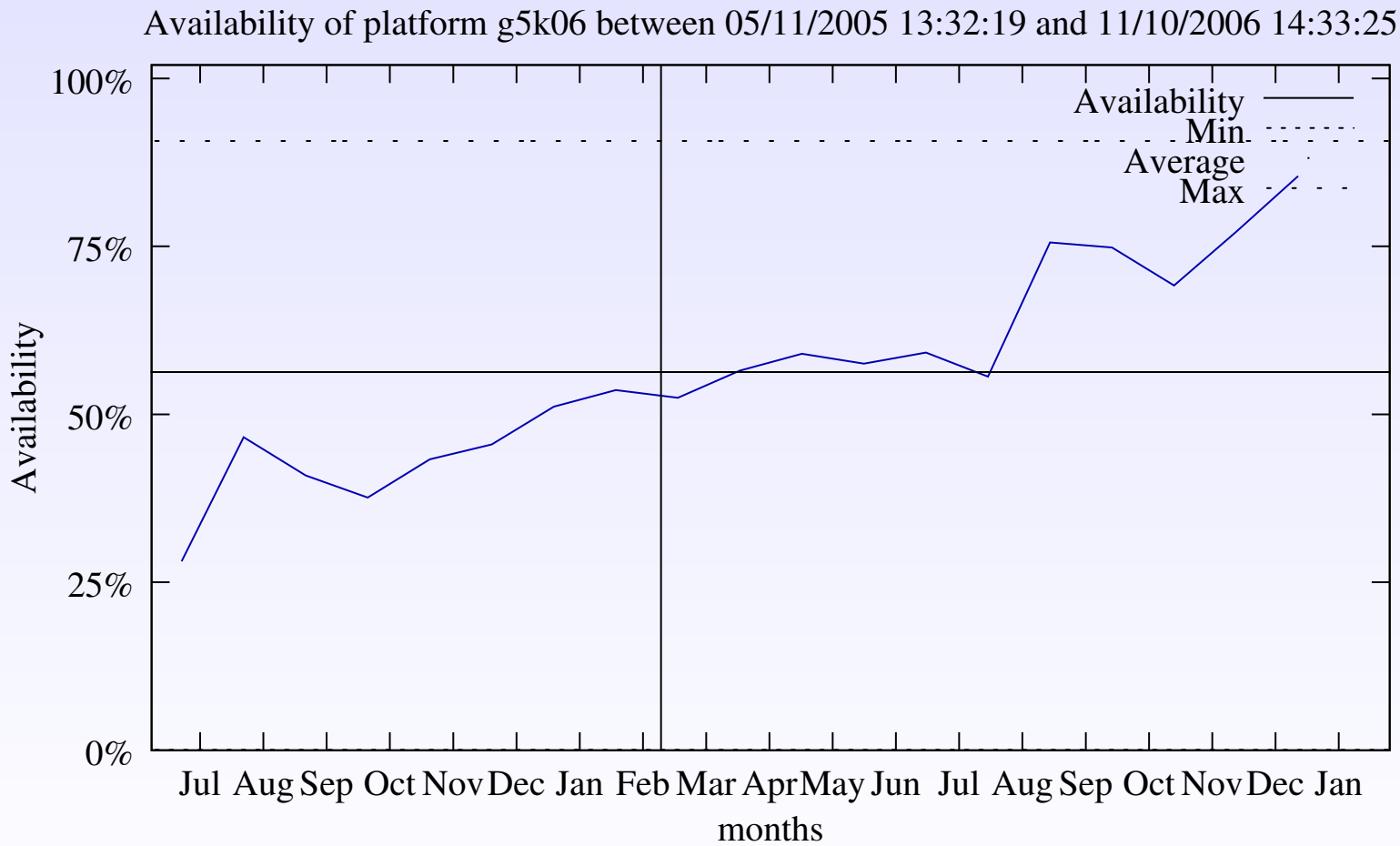
Future Work (III)

- Research of self-similarity factors in global availability
- Computation of Hurst parameter
- Useful for generating traces over long durations



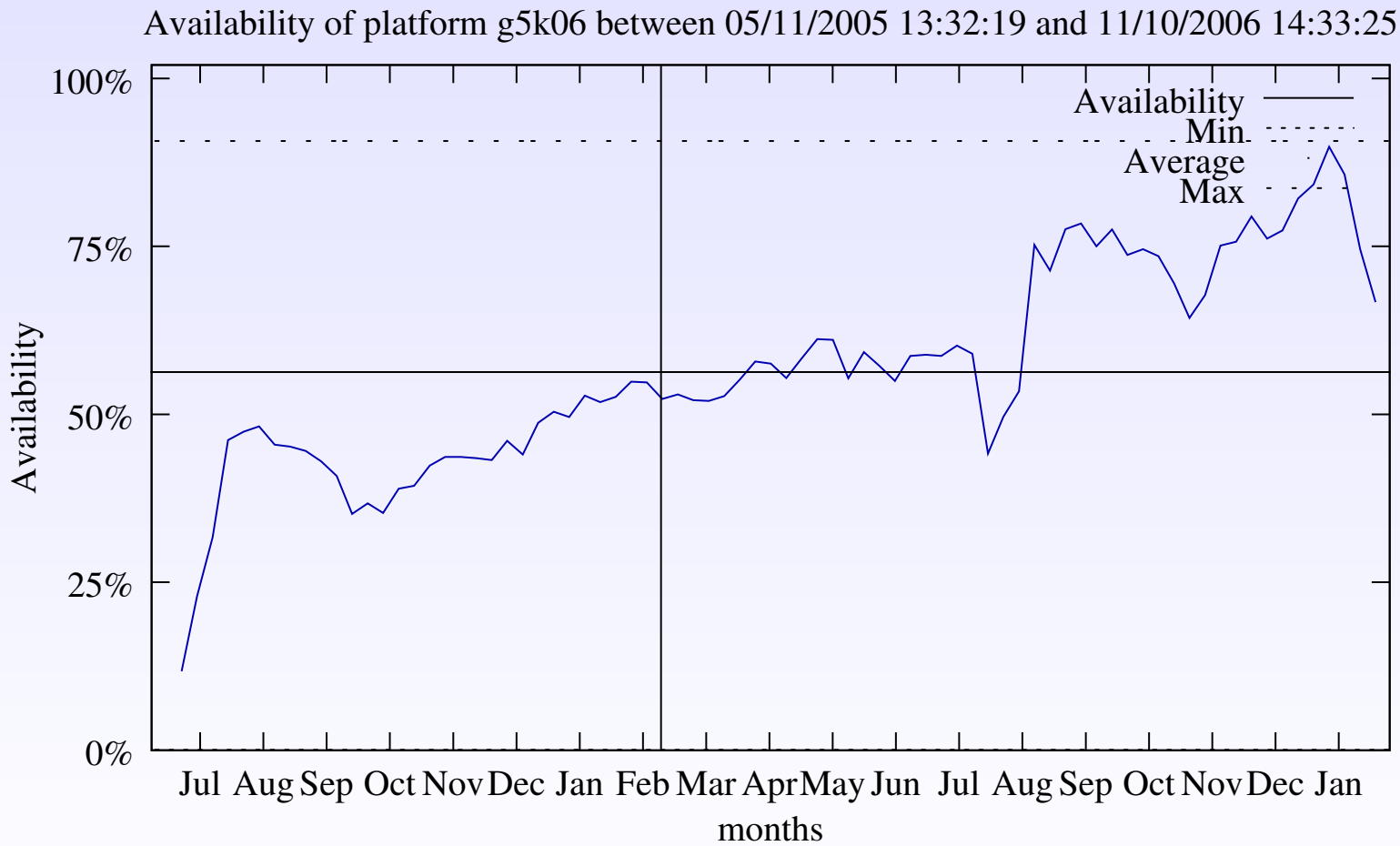
Future Work (IV)

- Digital signal processing techniques
- Decompose the graph into harmonics



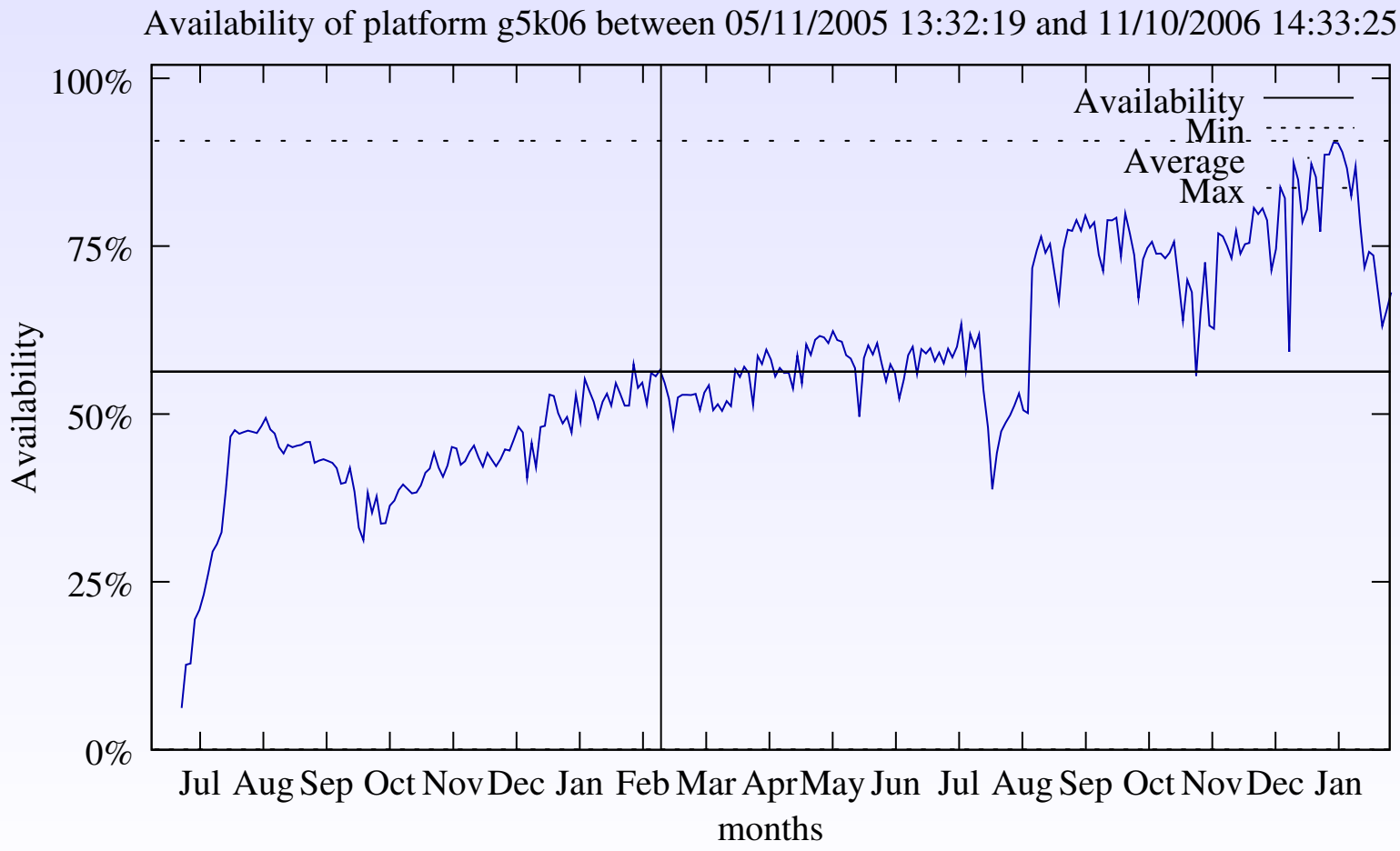
Future Work (IV)

- Digital signal processing techniques
- Decompose the graph into harmonics



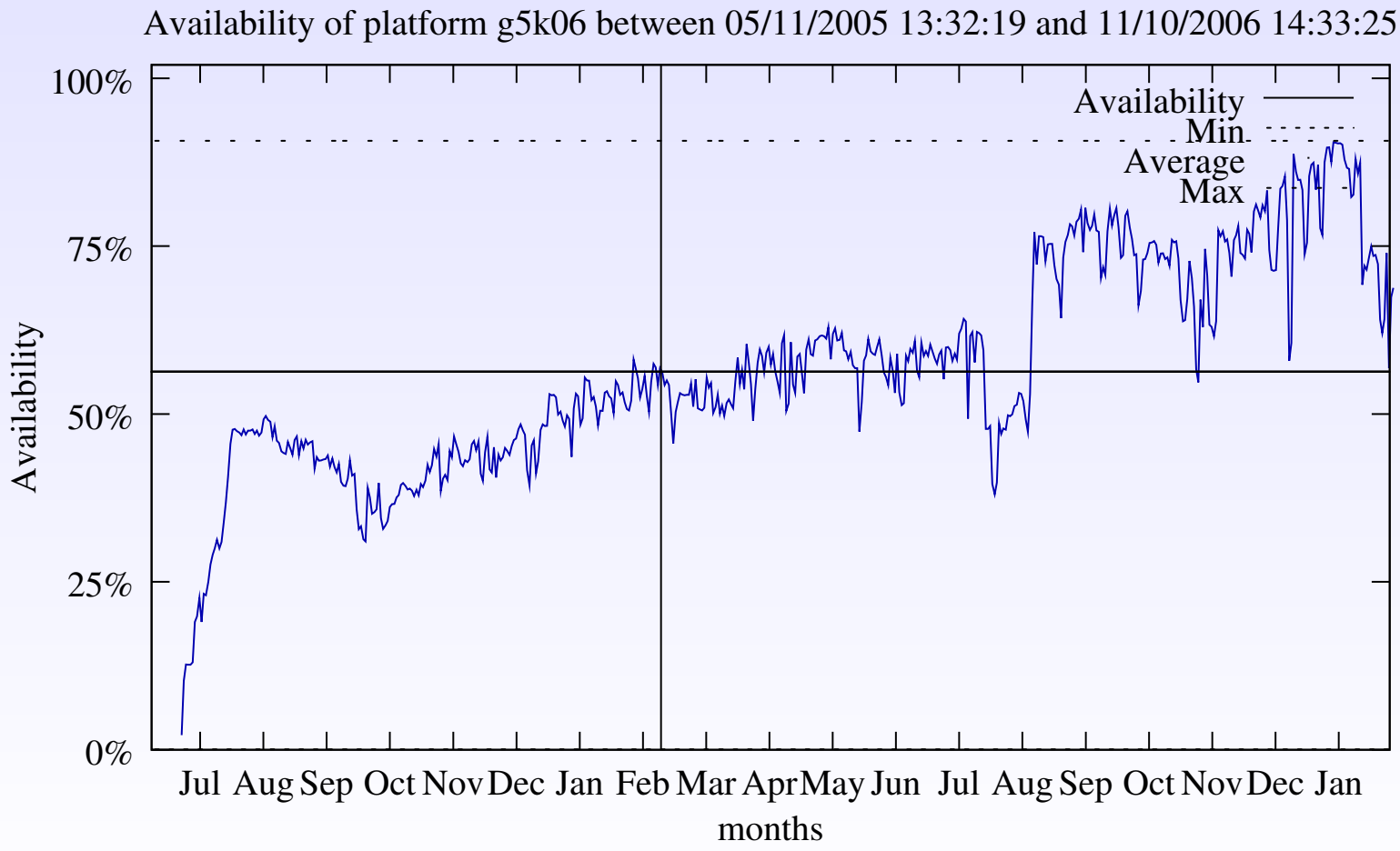
Future Work (IV)

- Digital signal processing techniques
- Decompose the graph into harmonics



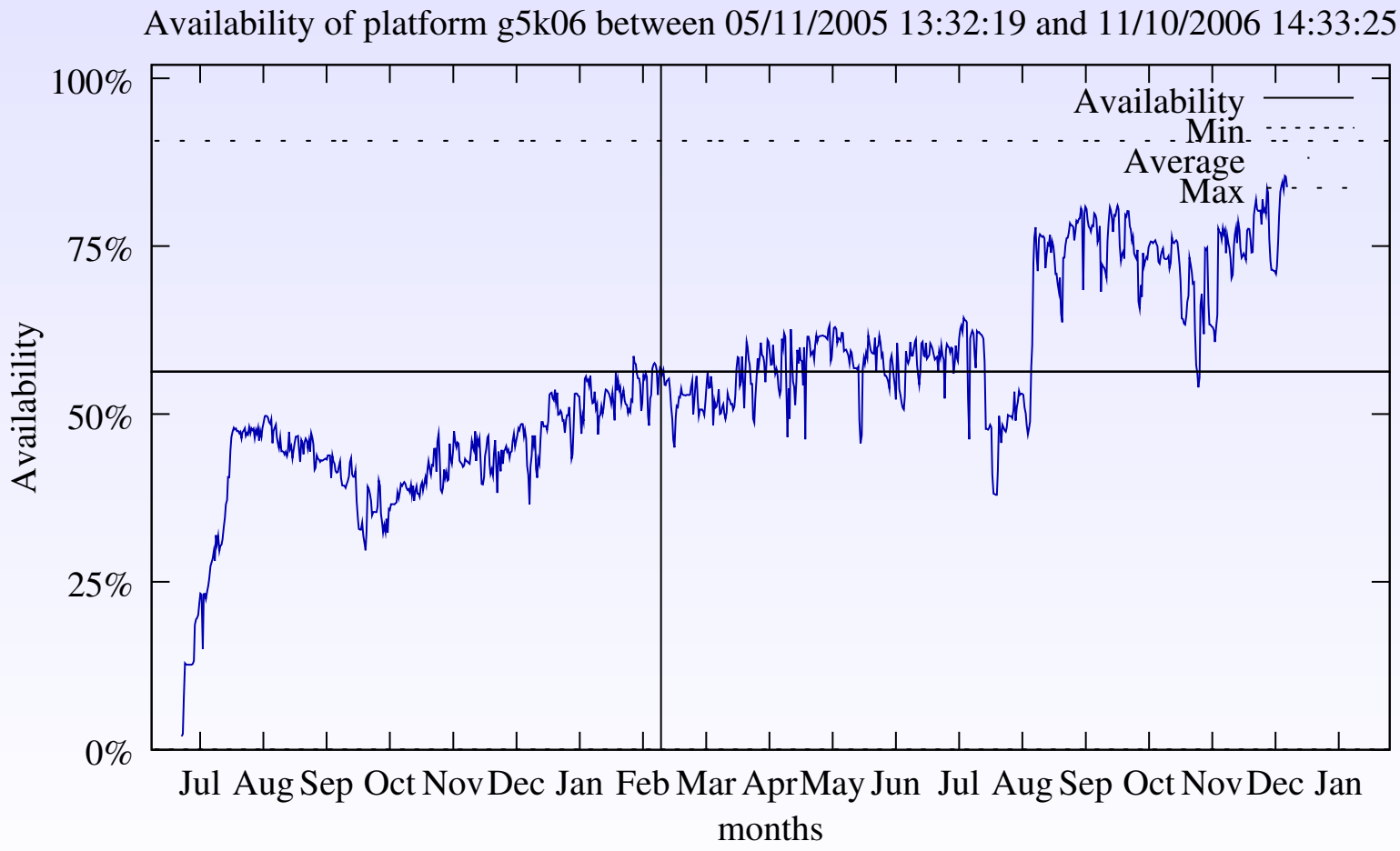
Future Work (IV)

- Digital signal processing techniques
- Decompose the graph into harmonics



Future Work (IV)

- Digital signal processing techniques
- Decompose the graph into harmonics



Conclusion

- Simple model, based on 4 components
- Components are well fitted by Log-Normal distributions
- Well-adapted to the design of fault-tolerant algorithms
- Lot of remaining work! :)