# Advanced System Monitoring in PTP

November 19, 2013 | Wolfgang Frings and Carsten Karbach



### Content

- PTP System Monitoring status quo
- 2 Short term enhancements
- 3 Long term enhancements

## Part I: System Monitoring – status quo

November 19, 2013 | Wolfgang Frings and Carsten Karbach



### **PTP Monitoring Scope**

- job and system monitoring of large-scale supercomputers
- examples for large-scale systems monitored with PTP:
  - BG/Q system JUQUEEN (JSC), LoadLeveler, 458K cores
  - Cray XT Jaguar (ORNL), TORQUE+ALPS, 299K cores
  - Kraken (NICS), Moab+PBS, 112K cores
  - Yellowstone (NCAR), LSF, 72K cores
- monitoring of multiple target systems in one perspective
- support for many batch systems (Grid Engine, LoadLeveler, Open MPI, PBS, Slurm, Torque, LSF)
- overview of the system on a single screen
- uniform interface to supercomputers
- based on monitoring application LLview



### **PTP Monitoring Perspective**

Elle Edit Navigate Search Broject Bun Feedback Sample Menu Services Window Help

Intervention         System Type           Access to law         System Type           Access to law System Type         System Type           Access to law System Type	atus Connec												
Bit         Description         Description           Active Adds Bit         Intractive Adds         Intractive					P	• 🔳 🖓 🗆	+ × °		iii syst	em: juqueen2.zai	n kfa-juelich de 😫		\$ \$
Active Jobs II       Inscription       Inscription <thinscription< th=""> <thinscription< th=""></thinscription<></thinscription<>	💩 Juquee	a -		System Typ	pa					R0000-M1	R0001-M1	R0002-M1	R0003-M1
Adve dos 20       Image: 1				IBM LoadLe	sveler (Blue Gene)								
Active das B       image das										R0000-M0	R0001-M0	R0002-M0	R0003-M0
Active das B       image das													
Affree dool 2       Image bound 2<													
Actor do B   Bacter dos										R0100-M1	R0101-M1	R0102-M1	R0103-M1
Actors dos 1         Encode dos                witer													
Actor do B   Bacter dos								-		B0100-M0	R0101-M0	R0102-M0	R0103-M0
present 1 zm k punk 64 407         000         1000 pt 14 m 0202 004 41         121 m 021 05 77         910         NAMAD           present 1 zm k punk 64 407         000         1000 pt 14 m 0202 004 41         121 m 021 05 77         910         NAMAD           present 1 zm k punk 64 407         000         1000 pt 14 m 0202 004 41         121 m 021 05 77         910         NAMAD           present 1 zm k punk 64 407         000         110 m 020 15 77         910         NAMAD         910         NAMAD           present 1 zm k punk 64 407         000         110 m 020 15 77         110 m 020 16 m 020 15 m 020 16	Active Jobs	88 III Inactive Jobs					×						
present 1 zm / Kapish é 420 m 1 Ma 202 (250 2 C 1 2 Ma 201 2 0 2 2 Ma 201 2 0 2 2 Ma 201 2 2 Ma 201 2 2 2 Ma 201 2 2 2 Ma 201 2 2 2 2 Ma 201 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2								<u> </u>					
preparent 1 zm k hunch 6, 400         6000         6000         11 % 2020         21 2020         11 % 2020         2200         11 % 2020         2000         6000         70000         7000         7000 <td< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>R0200-M1</td><td>R0201-M1</td><td>R0202-M1</td><td>R0203-M1</td></td<>										R0200-M1	R0201-M1	R0202-M1	R0203-M1
present 12 am Ka push e 450° (not)         2020 h 11 No 202 2102 cl 2 (2 lin 2010 6:00 m)         Bits 2010 h 200 h													
mpsereit 1 zm Meijneh 64 571 (m01         2200 m H No 2022 (104 Cl 2 U N 2016 303 m H No 2022 (104 Cl 2 U N 2016 303 m H No 2022 (104 Cl 2 U N 2016 303 m H No 2022 (104 Cl 2 U N 2016 303 m H No 2022 (104 Cl 2 U N 2016 303 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2022 (104 Cl 2 U N 2016 301 m H No 2016 (104 m H No 2012 (104 m H No 201 m H No 201 m H No 2012 (104 Cl 2 U N 2016 301 m H No 2012 (104 Cl 2 U N 2016 301 m H No 2010 (104 m H No 201 m H No 20										000000.000	000001 140		
personel 12 am Ke junch es 621°, color         22200 m 11 No 2022 v0 0° C (2 100 m 2) 0° 27 0° m (2 100 m 2) 0° 77 0° m (1 00 m 2) 0° 77 0° m (								111		R0200-M0	R0201-M0	180202-M0	R0203-M0
present 1 zm k lapin ha 647 (not 2000)         2200)         1 H No 200 2 11 2 12 (1 2 H No 200 2 17 42 (2 H No 200 2 17 42 (2 H No 200 2 1 42 (2 H No 200 1 2 H No 200 2 1 42 (2 H No 200 1 2 H No 200 1 1 H No 200 2 H (1 2 H No 200 1 1 H No 200 1 4 H No 200 2 H (1 2 H No 200 1 1 H No 200 1 4 H No 200 1 1 H No 200 1 4 H No 200 1 1 H No 200 1 4 (2 H No 200 1 1 H No 200 1 H No 200 1 1 H No 200 1													
payenti 1 am Mayah 6 457 (04)         6500 (n) 1 to 3212 02 24 G (2) 1 bits 2010 02.54         208         RARRIE           payenti 1 am Mayah 6 457 (04)         6500 (n) 1 to 3212 02 24 G (2) 1 bits 2010 02.54         208         RARRIE           payenti 1 am Mayah 6 457 (04)         6400 (n) 1 to 3212 02 24 G (2) 1 bits 2010 02.54         208         RARRIE           payenti 1 am Mayah 6 457 (04)         6400 (n) 1 bits 2010 02 14 (2) 12 bits 2010 05.55         208         RARRIE           payenti 1 am Mayah 6 457 (04)         6400 (n) 1 bits 2010 02 14 (2) (2) 12 bits 2010 05.55         208         RARRIE           payenti 1 am Mayah 6 457 (04)         6400 (n) 1 bits 2010 02 10 (2) bits 2010 (2) (2) (2) (2) (2) (2) (2) (2) (2) (2)								111		00000.M1	IP0301-M1	IP0302.M1	D0303-M1
parent 12 mit Mayerin 6, 4570 of 01         4790 h 11 M 2012 00 221 4 G 12 W 2012 00 226         208         RANNER           parent 12 mit Mayerin 6, 4570 of 02         4790 h 11 M 2012 00 221 4 G 12 W 2012 00 226         208         RANNER         208         RANNER           parent 12 mit Mayerin 6, 4570 of 04         4490 h 11 M 2012 00 12 0 G 12 W 2012 01 230         208         RANNER         208         RANNER           parent 12 mit Mayerin 6, 4570 of 04         4490 h 11 M 2012 00 12 0 G 12 W 2012 01 33         208         RANNER         208         RANNER           parent 12 mit Mayerin 6, 4570 of 04         4490 h 11 M 2012 00 12 0 G 12 W 2012 01 33         208         RANNER         208         RANNER           parent 12 mit Mayerin 6, 4570 of 04         4490 h 11 M 2012 01 20 G 12 W 2012 01 33         208         RANNER         208         RANNER           parent 12 mit Mayerin 6, 4570 of 04         4400 h 11 M 2012 01 30 G 12 W 2012 01 30         208         RANNER         2000 h 11 M 2002 01 M 200 h 11 M 2002 01 M 200 h 11 M 2002 M								11					
■         ■         Control         Accol         Acco								111					
Image: Section 1         Section 2         Control         Part in the Section 2         Control         Part in the Section 2         Control         Part in the Section 2         Part in the Sec								11		R0300-M0	R0301-M0	R0302-M0	R0303-M0
Nextages III         Conside Links on Links	jugusen1	l c1.zam.kfa-juelich.de.45740 n	004 43800	on 12 Nov 2012 01:39:01 C	12 Nov 2012 01:39:33	2048	RUNNING	ш					
Messages II         Conside         Remute Environments         Image: International Construction           ary         Value         Image: International Construction         Image: Internation         Image: International Construltion         <	iumaant	of ram Mainalich da 15751 n	00.4 43800	un 17 New 2012 05:19:37 C	12 Nov 2012 05:21:55	2048					Yman	Non-ten	les un
Markade         Markade         Production         Prodoct Add         Production         Production </td <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td><u>د</u></td> <td>P</td> <td>R0400-M1</td> <td>N0401-M1</td> <td>R0402-M1</td> <td>N0403-M1</td>								<u>د</u>	P	R0400-M1	N0401-M1	R0402-M1	N0403-M1
yy         Value         Value           yurdin         12/11/10/2002/016         12/11/10/2002/016           yurdin	Messages 1	🛛 😅 Console 🛛 🔒 Remote	Environments				~						
y parties Water Market	av	Value						18		R0400-M0	R0401-M0	R0402-M0	R0403-M0
1, пос. уп. (2) 1, пос. (2) 1		LL12111200250216						11					
Late: del Barray         US           Lata: del Barray         COS         Del Barray	g_shape_alloc	2x4x4x2x2								Contrast Mar	Inorrow we	Inorrow Mil	loocoo uu
g. utak         Ranning         Section 40         Section 40 <td>g_size_alloc</td> <td>128</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>P</td> <td>R0500-MI</td> <td>RUSUT-MT</td> <td>R0502-M1</td> <td>R0503-M1</td>	g_size_alloc	128							P	R0500-MI	RUSUT-MT	R0502-M1	R0503-M1
Integration         NO           Integration         NO         1	g_size_req	128											
Inspectidate         Mon 12 Nov 2012 00 25 40 CET           Non         Mon 12 Nov 2012 00 25 40 CET           Non         Mon 12 Nov 2012 00 25 40 CET           Non         Mon 12 Nov 2012 00 25 40 CET           Non         Mon 12 Nov 2012 00 25 40 CET           Non         Mon 12 Nov 2012 00 25 40 CET           Non         Mon 12 Nov 2012 00 25 40 CET           Non         Mon 12 Nov 2012 00 25 40 CET           Non         Nov 2014 00 PROVING           Nov 2014 00 PROVING         Nov 2014 00 PROVING										R0500-M0	R0501-M0	R0502-M0	R0583-M0
second         No         PORO MI         PORO													
τοφ         ρ457           1         0004.00         0001.40         0003.40         0003.40           κ         0454.606.9         0004.40         0003.40         0003.40			ET										
ame R494805q									REL	R0600-M1	R0601-M1	R0602-M1	R0603-M1
ame R49e48136q PC601-M0 PC602-M0 PC603-M0 PC603-M0													
								11		R0500-M0	R0601-M0	R0602-M0	R0603-M0
	ame odelist		0 00101 M0 MIC	00104 M0 M14				11					

#### November 19, 2013

#### Wolfgang Frings and Carsten Karbach



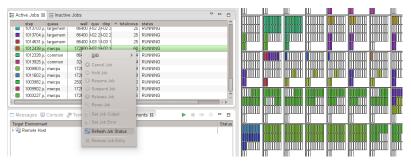
### **Monitoring Views**

- Nodes View renders target system architecture, maps jobs to compute resources
- Active Jobs View lists running jobs
- Inactive Jobs View lists queued jobs
- Monitoring View selects active target system, starts/stops monitoring
- Message View shows message of the day



### **User interaction**

- job management: cancel, get output/error
- filtering: show only user jobs, flexible filtering dialog
- linking information: click on job → highlight its nodes in Nodes View, show detail information in Message View
- change level of detail



### Part II: Short term enhancements

November 19, 2013 | Wolfgang Frings and Carsten Karbach



### **Job selection**

- raised in bug 403060
- allow selection of multiple jobs
- keep selected job selected until it is de-selected
- mark entire connected area of each job

Status Connection Name Configuration Name								in an	hannan	100000000	0000000	10000000	100000000	Terrer		
	local				Oper	1 MPI-Ge	neric-Ir	teractive								
<b>\$</b> 5	trestles				edu.	sdsc.tres	tles.ton	que.batch								
A	tive Jobs	22							~						imm	imm
	step	owner	queue	wall	queuedate		dispate	totak status				i i i i i i i i	mmm	MINIM	mmm	
	1305	cipres	shared	640800	2013-03-08 09	:00:26	201	8 RUNNING		LILLIUL .	цицици	LUILUILU	HUTT		,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	
Г	1305	hyowon	normal	172800	2013-03-08 14	:35:53	201	32 RUNNING								
	1305	depal	shared	432000	2013-03-09 05	:43:37	201	8 RUNNING		hhimmin	mmm	indirinini d	. Harrison Harrison	internet	hhimmin	
	1305	depal	shared	432000	2013-03-09 05	:43:55	201	8 RUNNING								
	1305	depal	shared	432000	2013-03-09 05	:44:05	201	8 RUNNING		0000000	innnnn	Immo	nannanna	In manage	Innonnon	Innanaa
	1305	depal	shared		2013-03-09 05		201	8 RUNNING			ասաս	ILLILLI		uuuu		
	1306	cipres	shared	604800	2013-03-09 13	:42:09	201	8 RUNNING								
	1306	seonah	shared	172800	2013-03-10 07	:21:55	201	16 RUNNING		1000000000	HUHHHHH	HUHHHH	22122122	Here was a second se	APRATA A	99049044
	1306		normal				201	32 RUNNING								
		mdejong	normal		2013-03-10 10		201	32 RUNNING		arnarnan	rmmn	Immin	<b>NUMBER</b>	mmm	innonnnn i	innannan
	1306		normal		2013-03-10 10		201	32 RUNNING				ງພາເພເບ	JUNDINU	JUUUUU	JULUUUUUUU	Juwiwi
	1307		normal	172800			201	32 RUNNING				10110110	01010100		Innonnon	1001001
	1307	hchen	normal	172800			201	32 RUNNING		uuuuu			ULLULLU		ULUULUUU	
		hchen	normal		2013-03-10 16		201	32 RUNNING								
	1307	hchen	normal	172800			201	32 RUNNING			<b>PROFILE</b>	EGHENNES	<b>Manager</b>	<b>HEREN</b>	aroaroan	
	1307	hchen	normal	172800			201	32 RUNNING								
	1307	jgraha8	normal	172800	2013-03-10 20	:55:26	201	256 RUNNING		100000000	i nonn hiddin	i na ka ka ka	minin	modifier	inno innih	1001001

Source: https://bugs.eclipse.org/bugs/attachment.cgi?id=228316



### Improved job localization

- adjustable minimum rectangle size
- currently set to 7 px
- ensure, that rectangle width/height are at least 7 px large, if possible take more space

Ele Edit Navigate Segrch Project Run Feedback Sample Menu Services Window Help																		
[전·포토] 십·西리 [월] [월] [월] [월] [월] · Q · Q, 네 G· 전 · 이 · 이 · 이 · 이 · 이 · · · · · · · ·														em Monitoring				
S Monitors 21	► = 2 + X = □	syst	em: judg	e 13											🗢 🗢 + 😐 t			
Status - Connection Name Juqueen Judge	Configuration Name IBM LoadLeveler (Blue Gene) de.fz-juelich.judge.torque.batch																	
Judge Juropa	de. fz-juelich. judge3. torque. batch de. fz-juelich. juropa. torque. batch																	
Active Jobs 12 III Inactive Jobs	v ⊨ □																	
- 1004472 inter 28800 8-27 1-8-27 1- 1004277 comm 86400 8-27 1:8-27 1: 1004275 comm 86400 8-27 1:8-27 1: 004275 comm 86400 8-27 1:8-27 1:	1 RUNNING 64 RUNNING 64 RUNNING		ļ	Ĵ.	<u>j</u>	į.	<u> </u>									Ī		
- 1004276. comm/86400 8-27 1:8-27 1: 1004061. comm/86400 8-27 1:8-27 1: 1004061. comm/86000 8-27 1:8-27 1: 1004061. comm/8000 8-27 1:8-27 1: 1004276. comm/86400 8-27 1:8-27 1: 1004261. comm/86400 8-27 1:8-27 1: 1004276. comm/86400 8-27 1:8-27 1:8-27 1: 1004276. comm/86400 8-27 1:	64 RUNNING			,									<b>4</b> 00000					
🗖 Messages 😫 Console 🛛 🍠 Terminal																		
No consoles to display at this time.	11 II V 11 V																	
		<														< >		



### Improved job localization

- adjustable minimum rectangle size
- currently set to 7 px
- ensure, that rectangle width/height are at least 7 px large, if possible take more space

Eile Edit Navigate Segrch Project Run	Feedback Sample Menu Services W	indow <u>H</u> elp			
	'   \$\$ ♥ O ♥ \$1 ♥ 60 ♥ 1 ♥ 61 ♥ 61		📣 Quick Access 📑	Resource 😨 C/C++ 🛗 Syste	am Monitoring
S Monitors 21	► = 2 + X = 0	🔛 system: judge 😫			+ = =
Status         Concertion Name           Judges         Judges           Judg	Centiguration Name EMI Last Centrel: (Bior Genn) de to jenich Judget Groupe Match de to jenich Judget Groupe Match de to jenich Judget Groupe Match Infalie: Mateus 1 RONNING del RO				
No consoles to display at this time.					



### Improved job localization

- adjustable minimum rectangle size
- currently set to 7 px
- ensure, that rectangle width/height are at least 7 px large, if possible take more space

Status         Connection Name         Configura           Juqueen         EM Load           Judge         de forgaria           Jurga         de forgaria           E Active Jobs 33         E inactive Jobs	> 9.      > .     > .     > .       ation Name	system: judge 12	cess 📄 📴 🕞 Resourc	te 🗟 C/C++ 🔛 System Monitoring
Status         Connection Name         Configura           Juqueen         BM Load           Judge         de fizze           Judge         de fizze           Jurepa         de fizze           Jurepa         de fizze           Et Active Jobs 131         Enactive Jobs	ation Name dLeveler (Blue Gene) fich.judge.torque.batch fich.judge3.torque.batch	system: judge 22		
Jungeen EN Laad Junge de Erjueit Junge de Erjueit Junge de Erjueit	dLeveler (Blue Gene) fich.judge.torque.batch fich.judge3.torque.batch			
			╾┸╾┸╾┘ <b>╠═╏═</b> ╏╾╹	
step         queue         vagat         daya         <	NNING NNING NNING NNING NNING NNING NNING			
No consoles to display at this time.	11 - V 11 V			



### **Provide customized LML layouts**

- get familiar with system architectures of supported target systems
- map system topology into LML layout
- great potential in customized layouts: level of detail, job filtering, showing node names

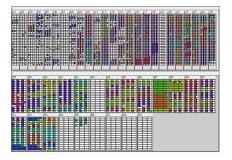
system :																 + • •
	<b>,</b>	<b>,</b>	<b>,</b>	1		ļ										
					<b>,</b>		ļ			ļ	, and the second se				ļ.	
															ļ	
						<b>H</b>	<b>H</b>		ļ.	<b>,</b>				 <b>,</b>	<b>,</b>	
										<b>P</b>						
		<b>,</b>	<b>.</b>				ļ	<b>J</b>						]		

define own grid



### **Provide customized LML layouts**

- get familiar with system architectures of supported target systems
- map system topology into LML layout
- great potential in customized layouts: level of detail, job filtering, showing node names



- hierarchy
- level of detail



### Job handling and additional job list

### Completed jobs

- jobs submitted externally disappear when completed
- if batch system does not list them, the job entries are lost
- idea: keep track of user's running jobs, which are removed on update

### New job list

- currently: active and inactive jobs
- better: submitted, active and completed jobs

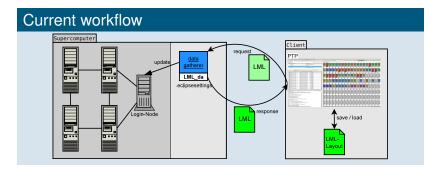
## Part III: Long term enhancements

November 19, 2013 | Wolfgang Frings and Carsten Karbach



### **Caching LML files**

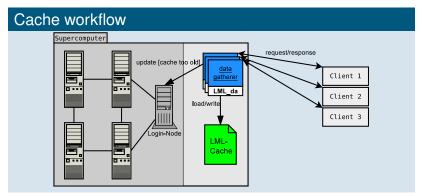
- multiple users on the same target system
- currently each user triggers separate LML\_DA workflow
- cache LML file in public directory (e.g. /tmp), use LML cache as data source





### **Caching LML files**

- multiple users on the same target system
- currently each user triggers separate LML\_DA workflow
- cache LML file in public directory (e.g. /tmp), use LML cache as data source



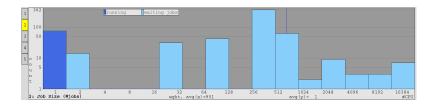


### New monitoring views

- derived from LLview, new monitoring types are possible
- fast overview on system statistics, history and prediction
- data description is already included in LML
- todo: data generation and visualization for new diagrams



### **Histograms**

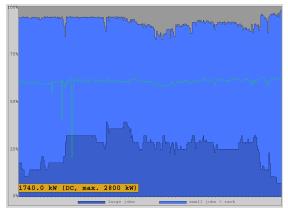


- statistical data rendered as histograms
- visualization of job parameter distribution: queue, size, waiting time



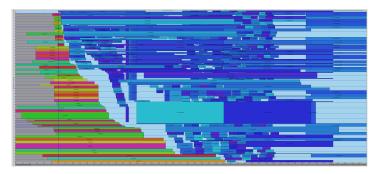
### Load history

- usage history of the target system (e.g. last 3 days)
- extendable for power/memory/accelerator usage
- requires LML log, switch to stateful server





### **Prediction diagram**



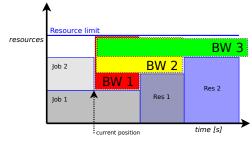
- Gantt chart visualization of future workload
- each rectangle represents one job, x-axis → time, y-axis → resources
- requires JuFo integration (see next slides) for simulation of future schedule

November 19, 2013



### JuFo – Overview

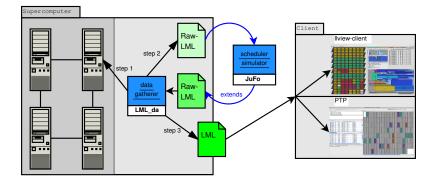
- configurable simulator for global job schedulers for on-line prediction of job dispatch dates
- based on analysis of JSC batch systems Moab and Loadleveler
- integrated with monitoring system LLview
- LML as configuration and communication data format
- use-cases:
  - user predicts start dates of submitted jobs
  - administrator simulates job scheduler performance with various input parameters, verifies scheduling rules



Wolfgang Frings and Carsten Karbach



### **JuFo integration**



- implemented in C++, additional installation step required
- simulation duration: 1-90 seconds  $\Rightarrow$  caching

November 19, 2013



### JuFo – Features

#### supported scheduling algorithms

- First-Come-First-Served
- List-Scheduling
- Backfilling

#### available simulation parameters

- generic job prioritization
- advanced reservations
- jobs can request CPUs, GPUs, memory
- nodesharing
- queue constraints
- test framework for evaluating JuFo's accuracy

## Part IV: Conclusion

November 19, 2013 | Wolfgang Frings and Carsten Karbach

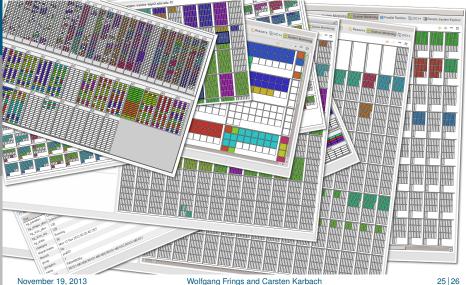


### Conclusion

- PTP provides monitoring system for large-scale supercomputers
- monitoring views: job lists, nodes view
- short term enhancements:
  - adjust job selection, multiple jobs
  - simpler detection of small jobs
  - create customized LML layouts
- Iong term enhancements:
  - LML file caching
  - new monitoring views: histograms, history, prediction
  - integration of JuFo



### Your ideas?





### Contact

E-mail:

c.karbach@fz-juelich.de, w.frings@fz-juelich.de

- LLview  $\rightarrow$  http://www.fz-juelich.de/jsc/llview
- $LML \rightarrow http://llview.zam.kfa-juelich.de/LML$