Analýza dat na gridu

Jiří Chudoba

11.9.2009

Fyzikální ústav AV ČR, v.v.i, Praha

Přehled

- Výpočetní infrastruktura pro ATLAS
- Testy STEP09
- HammerCloud testy

Zpracování dat pro ATLAS

- Tier0 = CERN
- 10x Tier1
  - úchova části RAW dat, ESD, AOD
  - reprocessing
- 67 Tier2
  - MC simulace
  - uživatelská analýza
- Tier3

Tier2 ve FZÚ - praguelcg2

- golias25 and ce1 – 2 CEs pro PBSPro server golias
  - lcgatlas pro všechny uživatele
  - lcgtlasprod omezeno pro speciální uživatele (production, pilot a sgm role)
  - SL4.7 WNs, 344 CPUs, 538 jader
    - Xeon 5420 @ 2.5 GHz se 4 jádry (8 HEP-SPEC06/jádro), Xeon 5160 @ 3 GHz se 2 jádry (10 HEP-SPEC06/jádro), Xeon@3.06 Ghz (4 HEP-SPEC06/jádro)

Letošní přírůstky

- ce2 a cream1 – 2 CEs pro Torgue server
  - gridatlas fronta pro ATLAS
  - SL5 WNs
    - 248 CPUs, 992 jader, jen Xeon E5420@2.5 GHz a E5440@2.83 GHz, 16 GB RAM
    - 8.4 a 8.7 HEP-SPEC06/jádro
  - dosud převážně pro D0 a ALICE
  - ATLAS podporuje (oficiálně) SL5 od srpna 2009

Diskový prostor - DPM

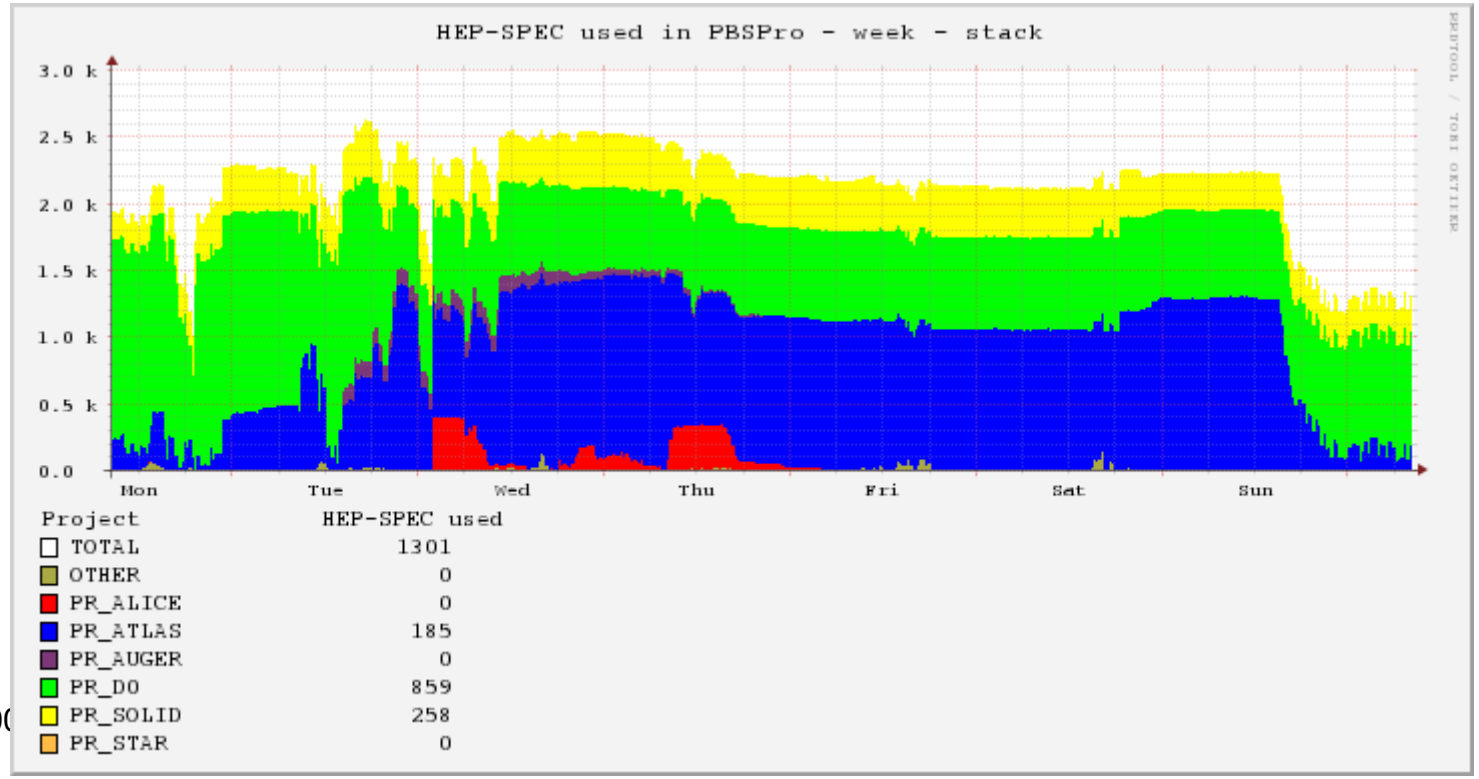
- 1 dpm head node golias100
- ATLAS tokeny definovány v poolu heppool1
  - 4 disk servery: se3 (goliasx98), se4, dpmpool1, dpmpool2
  - celkem 90 TB, volno 50 TB

Space-Tokens at Tier2s

Site	Space Token (Available / Used in TB)					
	DATA	MC	PROD	GROUP	LOCALGRP	SCRATCH
CSCS-LCG2	27 / 11	27 / 24	4.6 / 0.5	-	4.9 / 1	4.8 / 2
CYFRONET-LCG2	61 / 6	20 / 19.5	5 / 0.5	5 / 0.1	5.1 / 0.2	2 / 2
DESY-HH	62 / 3	82 / 76	4 / 0.1	3 / 0	10.7 / 10.4	5.7 / 5.2
DESY-ZN	69 / 8	69 / 61	4.7 / 0.1	1.8 / 1.8	9.3 / 1.5	1.8 / 1.8
GOEGRID	27 / 13	27 / 17	4.6 / 0.5	4.6 / 0	9 / 8	1.8 / 0.7
HEPHY-UIBK	15 / 2	12 / 10	3 / 0.1	1 / 0	1 / 0	3 / 0.2
LRZ-LMU	24 / 5	52 / 37	8.6 / 0.3	8 / 0	19 / 13	8 / 2
MPPMU	18.7 / 3.4	23 / 17	1.9 / 0.1	3.7 / 0	18 / 12	13.6 / 0.4
PRAGUELCG2	10 / 1.7	20 / 10	5 / 0.1	5 / 0	2 / 0.3	1 / 0.3
PSNC	1 / 0.1	1 / 0.4	5 / 0.1	-	-	1 / 0
UNI-DORTMUND	0.1 / 0	1.9 / 0.8	5 / 0.1	-	0.9 / 0	0.9 / 0.1
UNI-FREIBURG	38 / 7	37 / 17	3 / 0.1	3 / 0	20 / 18.5	5 / 0.3
UNI-SIEGEN	1 / 0	-	0.5 / 0	-	-	0.5 / 0
WUPPERTAL	21 / 1.4	20 / 15	4 / 0.4	8 / 0	0.5 / 0	7 / 5

MOU pro 2009

- 2009: 1504 HEP-SPEC06, 72 TB
  - **ATLAS: 624 HEP-SPEC06, 37 TB**
    - 78 cores (8 HEP-SPEC06/core)
- 2010: 2548 HEP-SPEC06, 201 TB – právě se upřesňuje

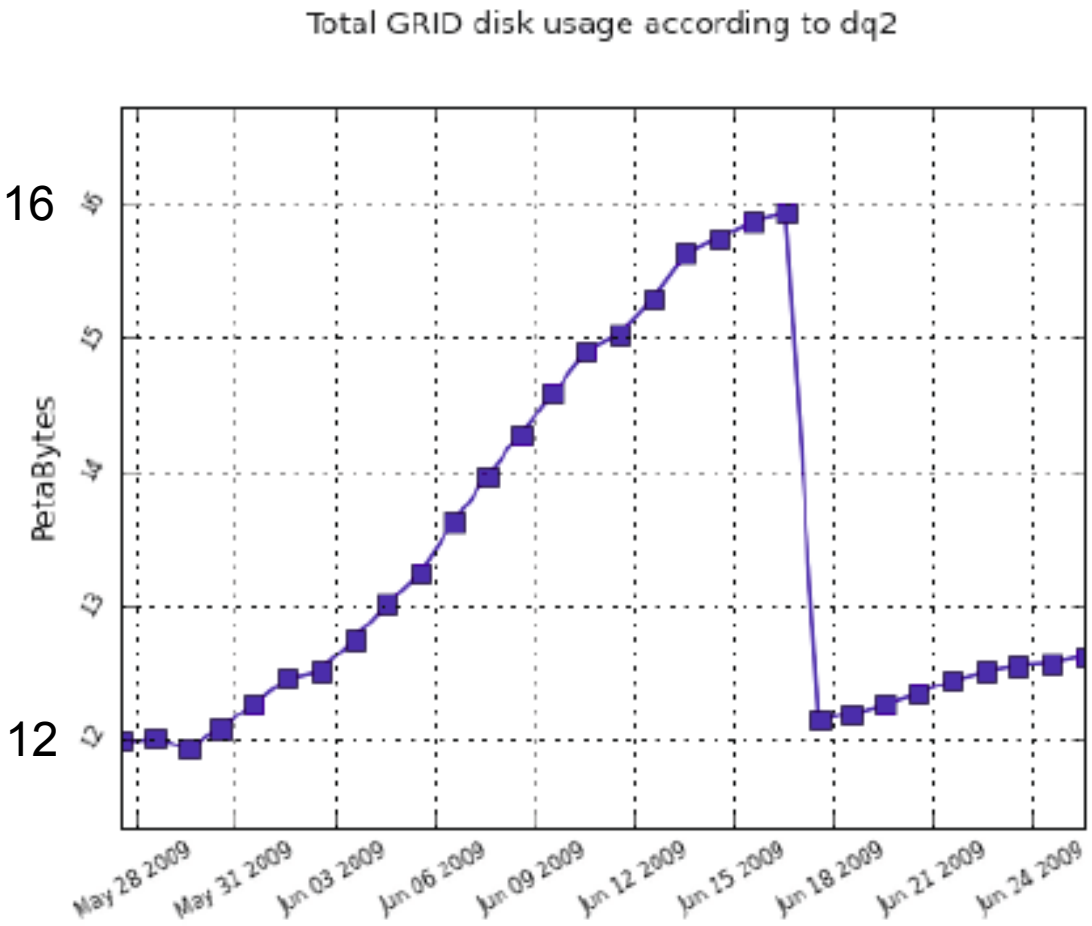


STEP09

- STEP09 = Scale Testing for the Experiment Program 09,
- Offline computing systems commissioning test
- Zahrnul všechny hlavní výpočetní aktivity
  - Monte Carlo simulace
  - Distribuce dat
  - Reprocessing v Tier-1s
  - Uživatelská analýza: Hammercloud
  - ATLAS Central Services Infrastructure
- testy probíhaly zároveň pro všechny LHC experimenty

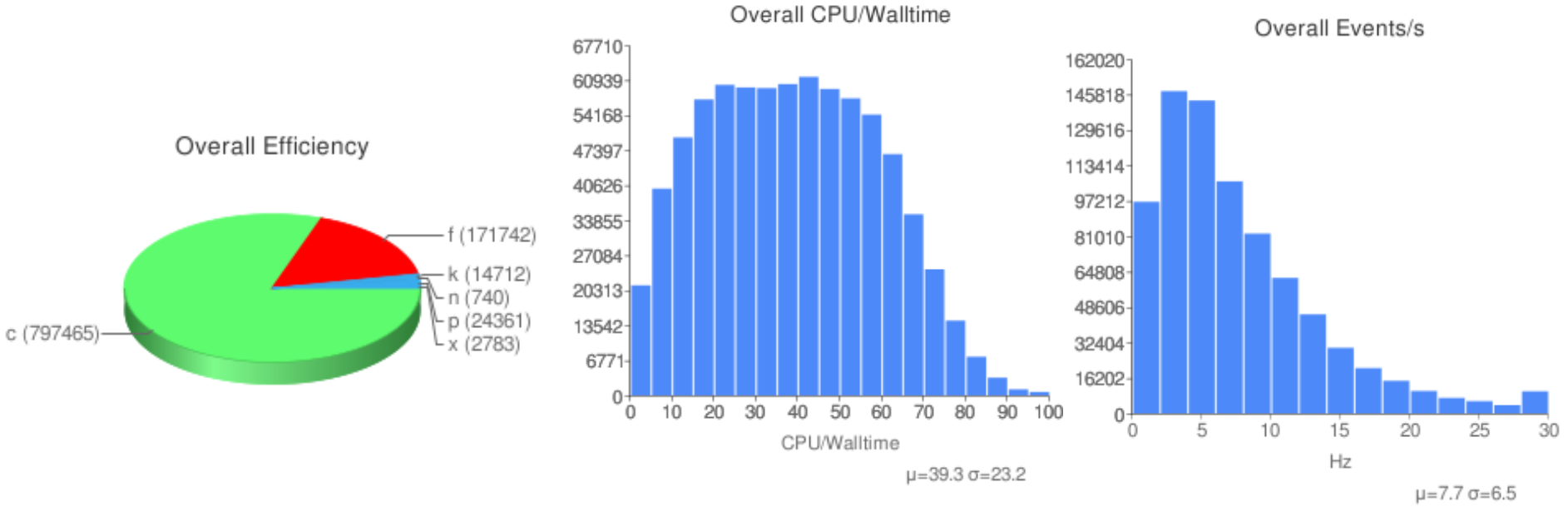
Distribuce dat

4 PB dat!



Hammercloud

- Simulace zatížení analyzačními úlohami
- 26 * 10^9 případů
- Mean Events/s = 7.7Hz
- Mean CPU/Walltime = 0.39



Celková eff: 82.3% Celkový výkon: 28.6kHz

STEP09 Cloud Results

Cloud	# Jobs	# Successful	# Failed	Efficiency	#files	#events	Hz	CPU/Wall
CA	41890	32306	9584	0.771	87117	757520054	4.2	34.4
DE	176076	135218	40858	0.768	557395	4991372555	7.9	40.8
ES	72562	62565	9997	0.862	236690	2150478621	10.0	44.3
FR	166427	144658	21769	0.869	557571	5050911395	9.3	44.5
IT	59163	52990	6173	0.896	311061	2798011153	6.5	32.6
NG	16730	14551	2179	0.870	20179	172708698	7.3	
NL	66632	37171	29461	0.558	154452	1352903529	7.9	35.0
TW	24178	19544	4634	0.808	86293	833817261	15.2	48.4
UK	181394	145222	36172	0.801	439084	3984651767	7.1	39.8
US	163004	153240	9764	0.940	465393	4169999722	6.0	33.9

STEP09 Merged Site Results

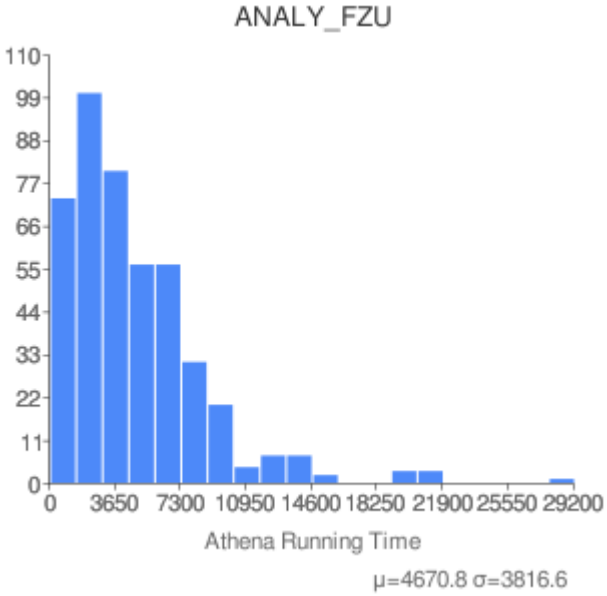
Cloud	Site	# Jobs	# Succ.	# Failed	Eff.	#events	data share	SE Size(TB)	#events/share	Hz	CPU/W
DE	WUPPERTALPROD	5783	1212	4571	0.210	19891044	0.17	63.0	117006141.176	4.7	29.7
DE	UNI-FREIBURG	8805	5320	3485	0.604	126367206	0.17	106.0	743336505.882	7.7	45.5
DE	PRAGUELCG2	870	256	614	0.294	3268970	0.05	23.0	65379400.0	0.7	8.9
DE	MPPMU	10641	5885	4756	0.553	270364769	0.16	89.0	1689779806.25	6.8	34.6
DE	LRZ-LMU	15814	11241	4573	0.711	278457039	0.16	146.0	1740356493.75	8.7	40.1
DE	HEPHY-UIBK	2056	960	1096	0.467	19368361	0.10	36.0	193683610.0	4.5	19.5
DE	GOEGRID	13798	10712	3086	0.776	673230762	0.15	76.0	4488205080.0	5.7	26.0
DE	FZK-LCG2	16286	11089	5197	0.681	424521314	1.00	-1.0	424521314.0	4.1	23.0
DE	DESY-ZN	26759	26147	612	0.977	1115755488	0.50	157.0	2231510976.0	11.3	56.7
DE	DESY-HH	18723	17629	1094	0.942	681135020	0.50	180.0	1362270040.0	9.3	43.9
DE	CYFRONET-LCG2	24271	16889	7382	0.696	386683806	0.17	99.0	2274610623.53	7.4	32.7
DE	CSCS-LCG2	29541	27125	2416	0.918	970253324	0.17	70.0	5707372494.12	6.7	43.1

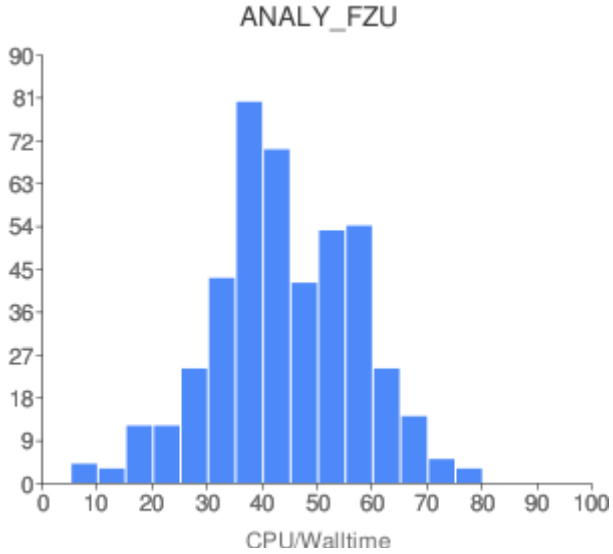
mnoho úloh ukončeno kvůli časovým limitům
sdílená linka 1 Gbps pro přístup k SE

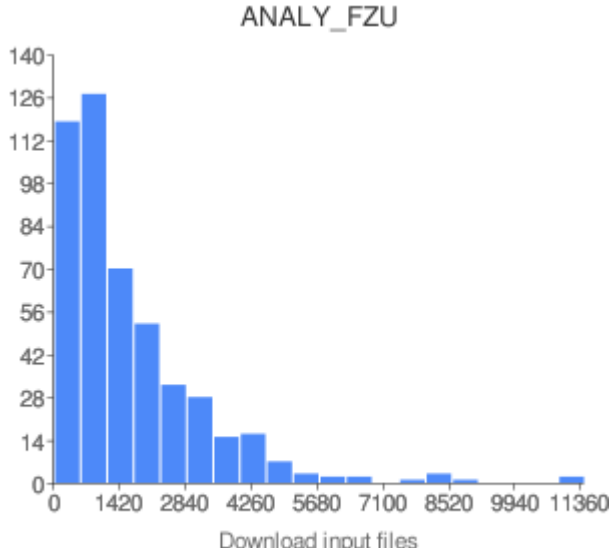
Opakování testů v DE 7/2009

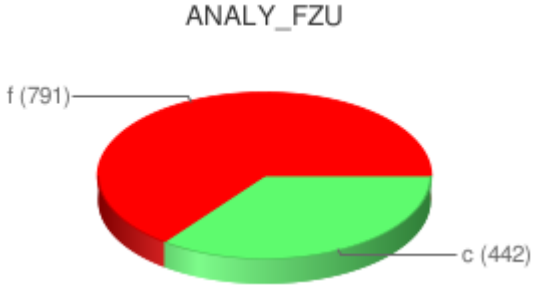
- 3 testy s různým přístupem k datům
  - panda – lcg-cp
  - dcap/rfio
  - FileStager - rfcp

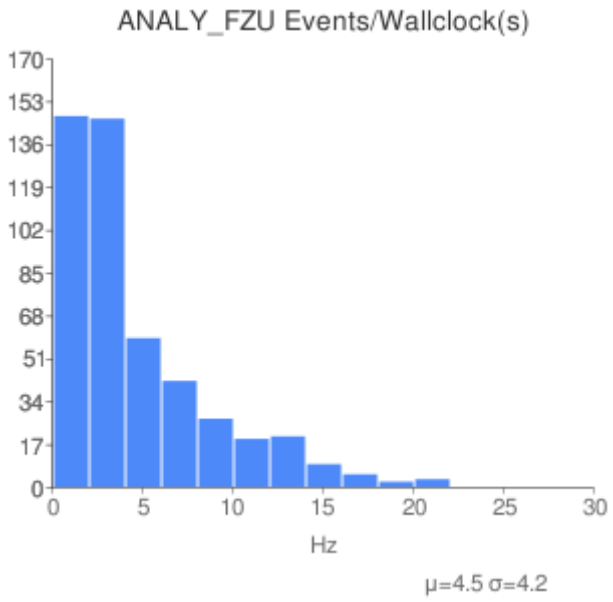
Test 525

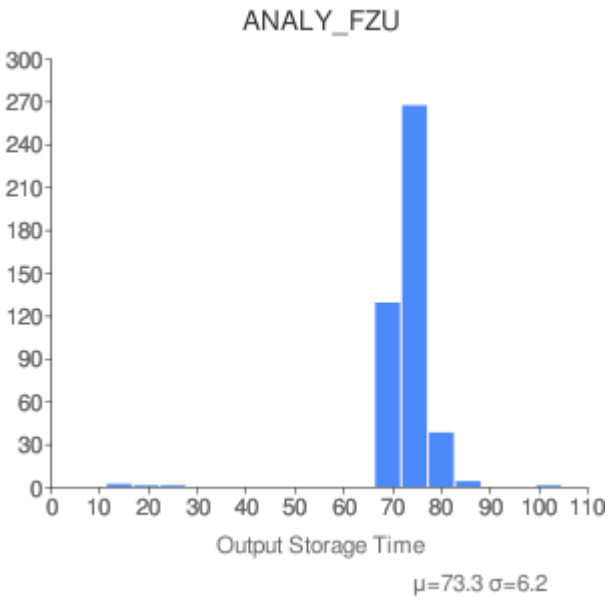

ANALY_FZU
110
99
88
77
66
55
44
33
22
11
0
0 3650 7300 10950 14600 18250 21900 25550 29200
Athena Running Time
μ=4670.8 σ=3816.6


ANALY_FZU
90
81
72
63
54
45
36
27
18
9
0
0 10 20 30 40 50 60 70 80 90 100
CPU/Walltime

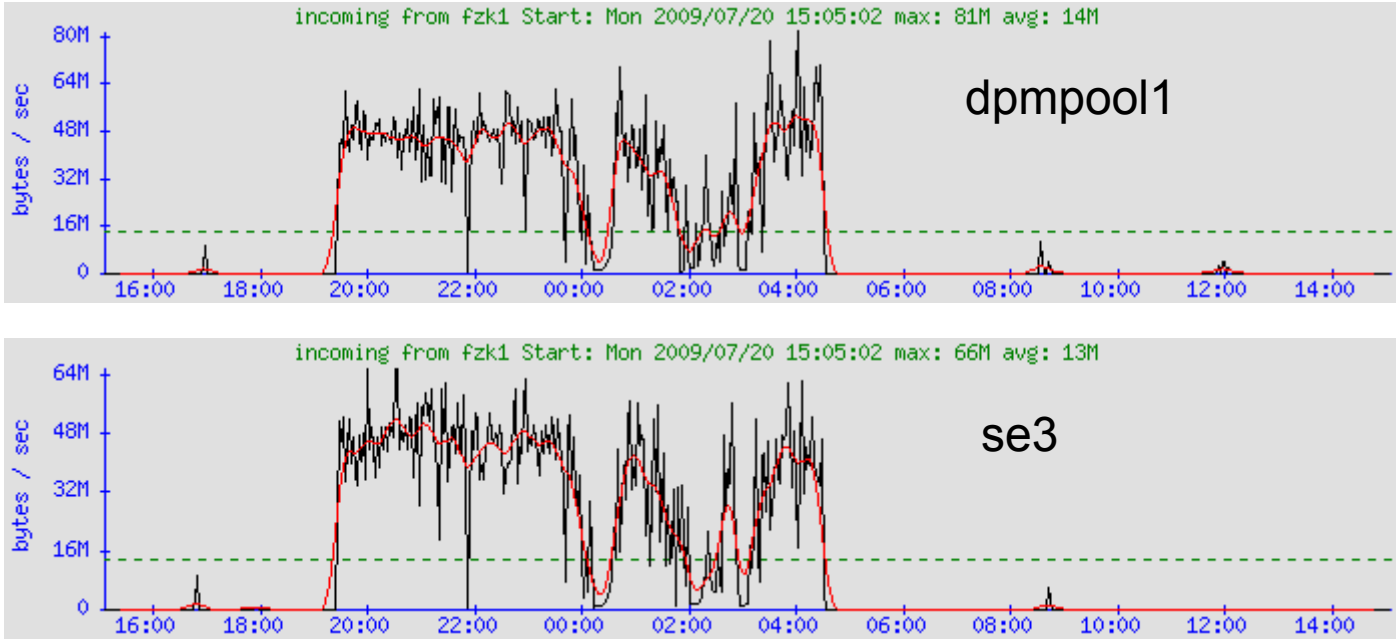

ANALY_FZU
140
126
112
98
84
70
56
42
28
14
0
0 1420 2840 4260 5680 7100 8520 9940 11360
Download input files


ANALY_FZU
f (791)
c (442)


ANALY_FZU Events/Wallclock(s)
170
153
136
119
102
85
68
51
34
17
0
0 5 10 15 20 25 30
Hz
μ=4.5 σ=4.2

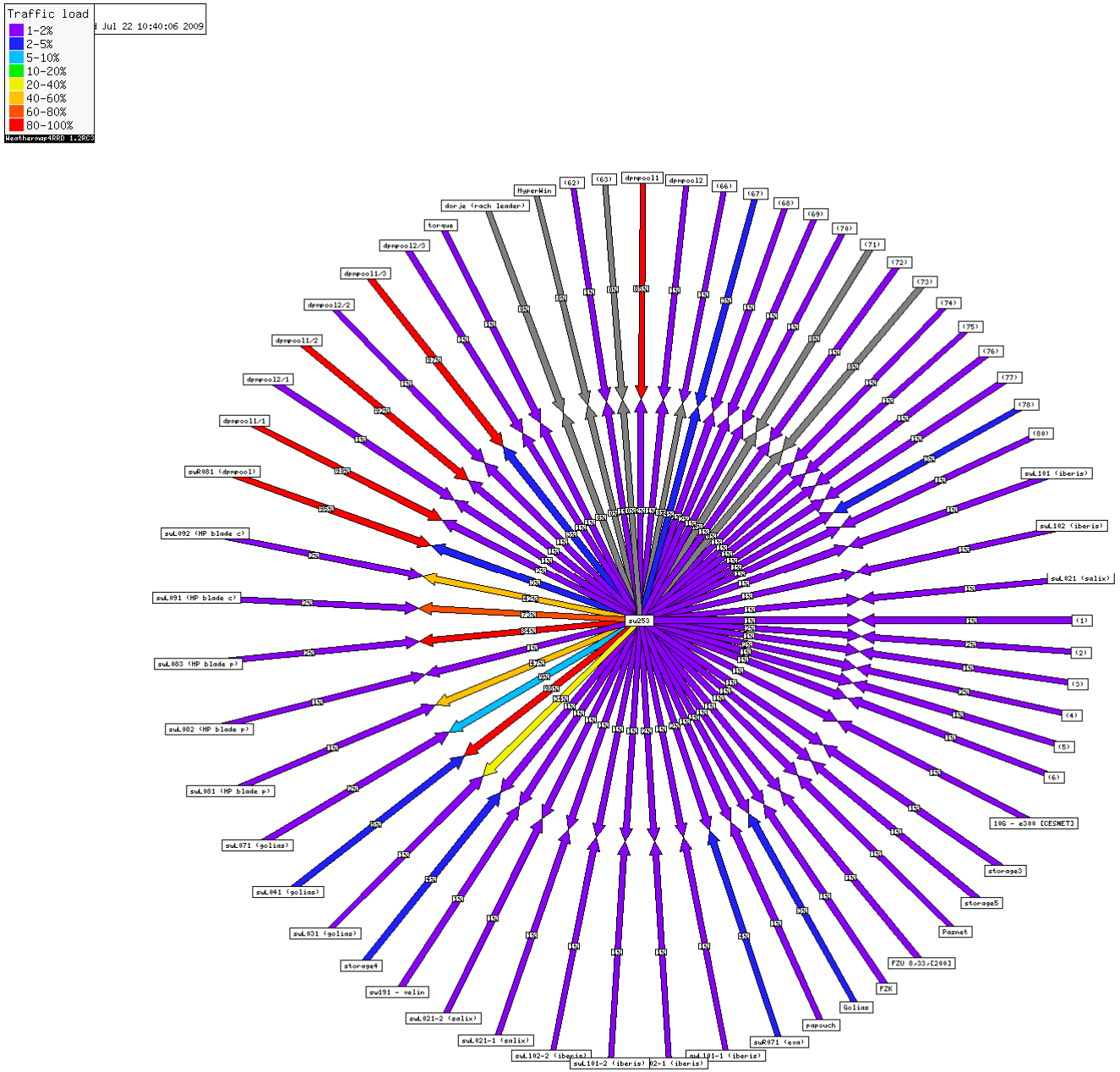

ANALY_FZU
300
270
240
210
180
150
120
90
60
30
0
0 10 20 30 40 50 60 70 80 90 100 110
Output Storage Time
μ=73.3 σ=6.2

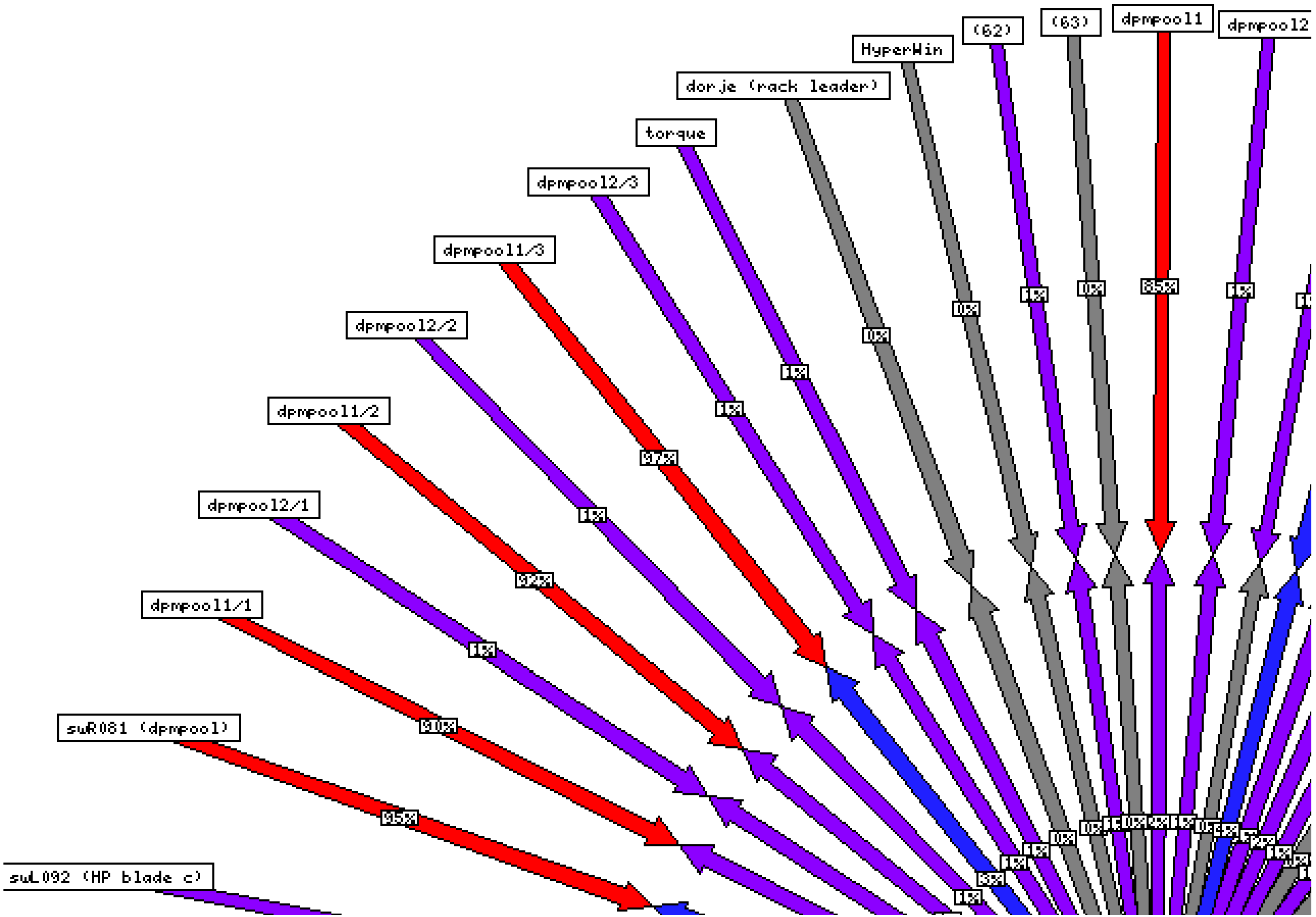
Test 525



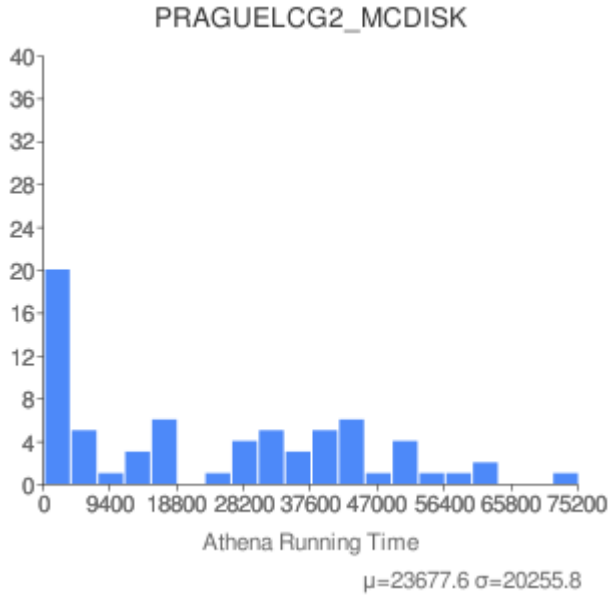
Transfer top DS z FZK začal ve stejný čas, 1 Gbps link saturována

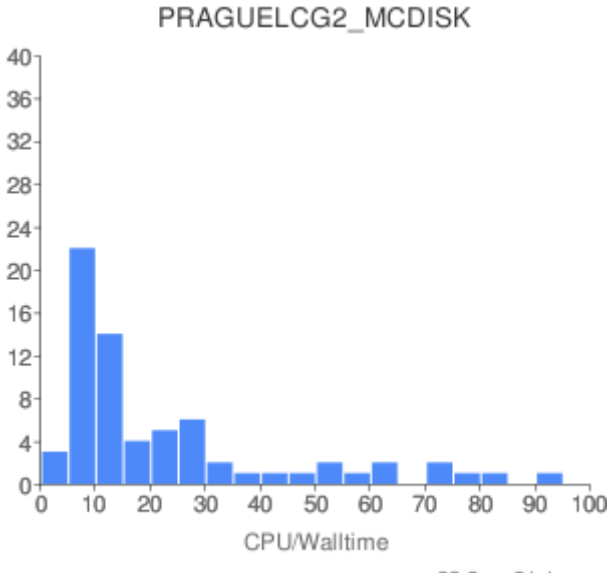
pád serveru se3 22.7., reboot, problémy s některými službami odstraněny až o několik hodin později

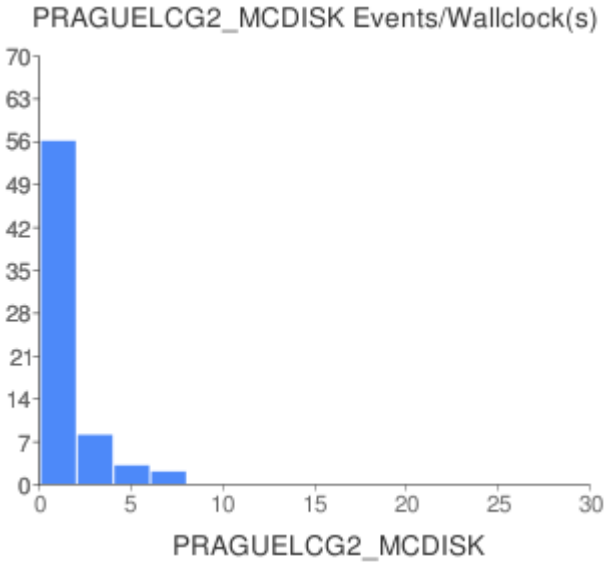

Traffic load
1-2%
2-5%
5-10%
10-20%
20-40%
40-60%
60-80%
80-100%
Jul 22 10:40:06 2009
dpmpool1
dpmpool2
HyperWin
dorje (rack leader)
torque
dpmpool2/3
dpmpool1/3
dpmpool2/2
dpmpool1/2
dpmpool2/1
dpmpool1/1
sw0253
sw191 - velin
storage4
storage3
storage5
Pasnet
FZK
Golias
papouch
10G - e300 [CESNET]

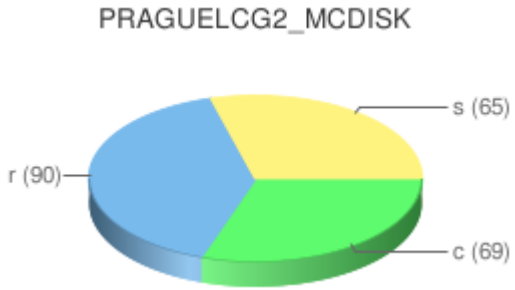

dpmpool1
dpmpool2
(62)
(63)
HyperWin
dorje (rack leader)
torque
dpmpool2/3
dpmpool1/3
dpmpool2/2
dpmpool1/2
dpmpool2/1
dpmpool1/1
suR081 (dpmpool)
suL092 (HP blade c)

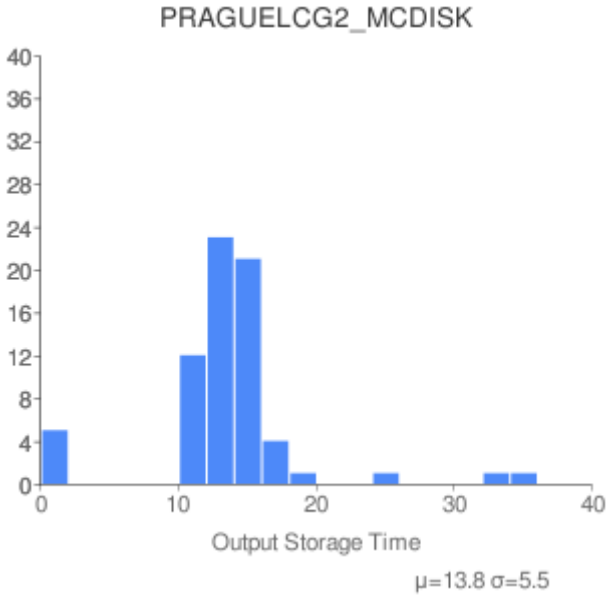
Test 531

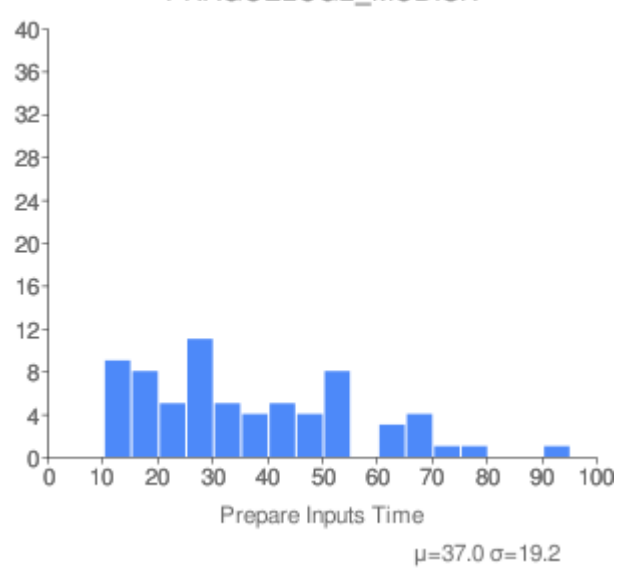

PRAGUELCG2_MCDISK
0 9400 18800 28200 37600 47000 56400 65800 75200
Athena Running Time
μ=23677.6 σ=20255.8

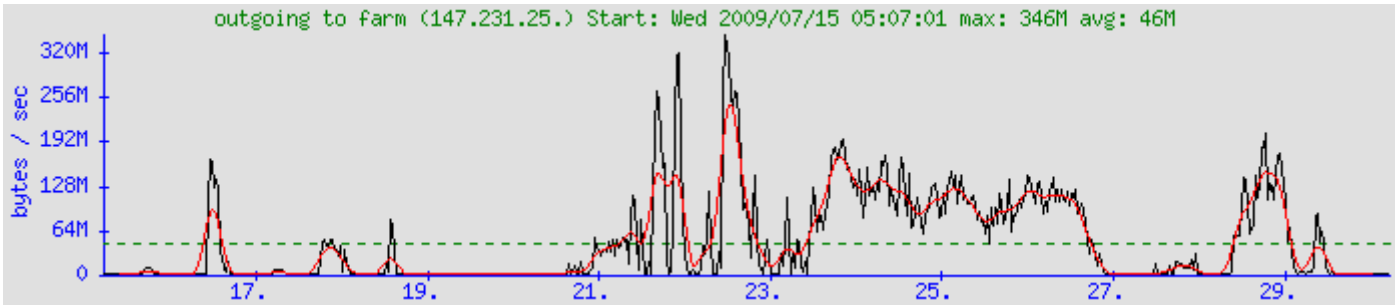

PRAGUELCG2_MCDISK
0 10 20 30 40 50 60 70 80 90 100
CPU/Walltime


PRAGUELCG2_MCDISK Events/Wallclock(s)
0 5 10 15 20 25 30


PRAGUELCG2_MCDISK
s (65)
r (90)
c (69)

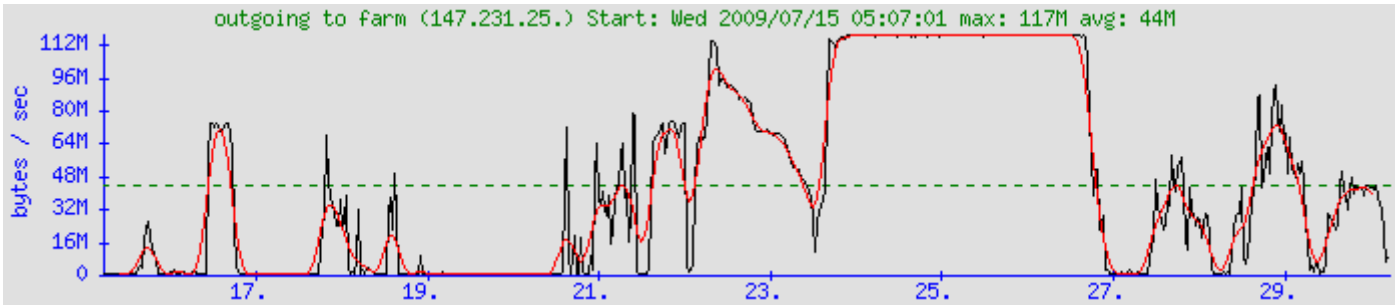

PRAGUELCG2_MCDISK
0 10 20 30 40
Output Storage Time
μ=13.8 σ=5.5


PRAGUELCG2_MCDISK
0 10 20 30 40 50 60 70 80 90 100
Prepare Inputs Time
μ=37.0 σ=19.2

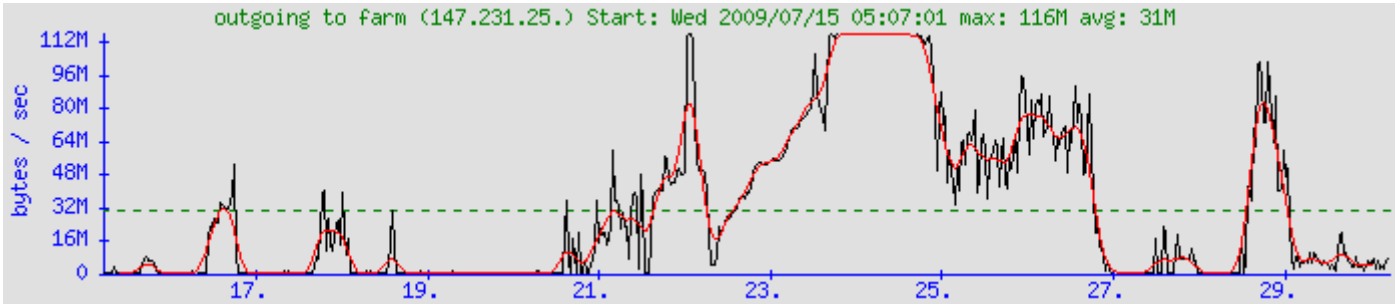


dpmpool1

3x1Gbps



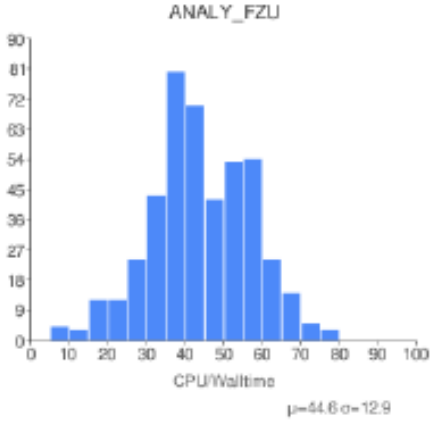
se3

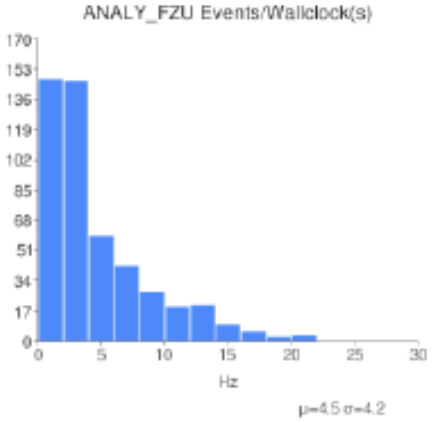


se4

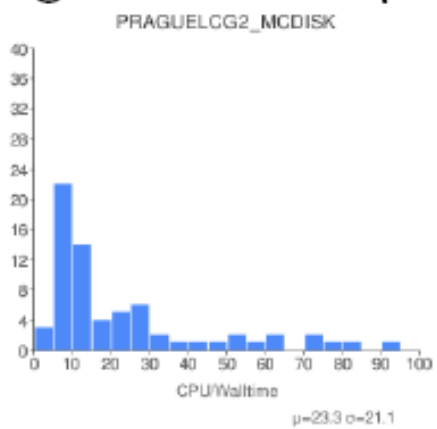
Recent HC test, FZU

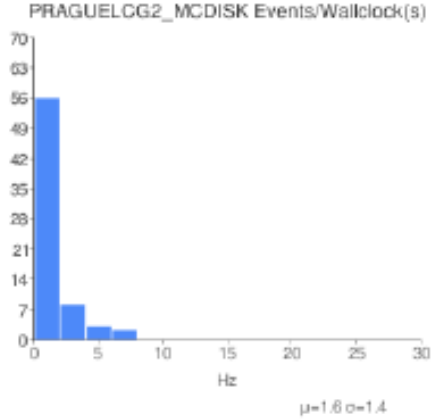
Panda



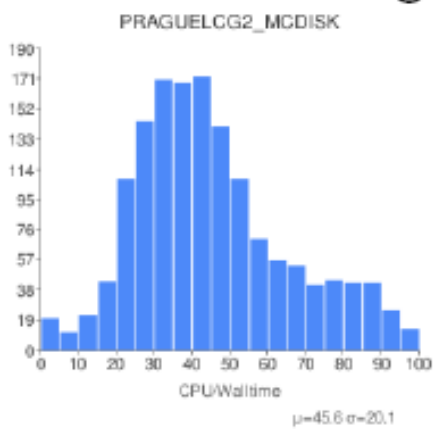


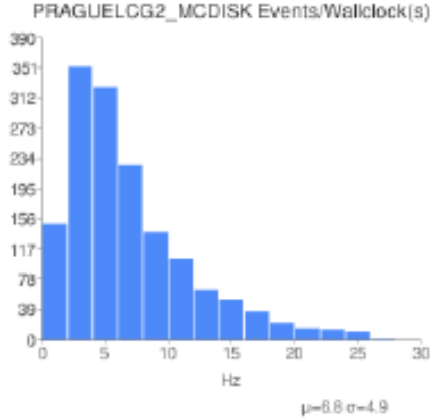
gliteWMS dcap





gliteWMS FileStager





	Jobs completed	Files	Events
Panda	442	1257	10595638
rfio	161	513	4156641
FileStager	1502	10066	88522808

Recent HC test - No. of Files

	Panda	dcap/rfio	FileStager
CSCS	10425	5543	14063
CYF	14962	241	0
DESY-HH	27529	3843	18892
DESY-ZN	13345	6419	9929
FZK	27662	2034	14231
GOEGRID	14164	6080	18287
HEPHY-UIBK	2303	1307	8
LRZ-LMU	21909	3321	12818
MPPMU	8050	5990	2995
PRAGUE	1257	513	10066
UNI-DORTMUND	470	15	26
UNI-FREIBURG	7786	5634	14533
WUPPERTALPROD	5732	152	140

There are still some bad sites despite some changes - we need to get ready for data very soon !

Recent HC test - Job failures

	Panda			dcap/rfio			FileStager		
	Failed	Compl.	F/(F+C)	Failed	Compl.	F/(F+C)	Failed	Compl.	F/(F+C)
CSCS	894	3155	22,1%	117	863	11,9%	2	2483	0,1%
CYF	153	5349	2,8%	15	159	8,6%	0	0	0,0%
DESY-HH	499	6613	7,0%	37	662	5,3%	0	2256	0,0%
DESY-ZN	1015	4589	18,1%	80	1046	7,1%	3	1467	0,2%
FZK	261	7447	3,4%	143	437	24,7%	6	2203	0,3%
GOEGRID	787	3588	18,0%	73	916	7,4%	71	2170	3,2%
HEPHY-UIBK	1359	818	62,4%	11	268	3,9%	3	6	33,3%
LRZ-LMU	256	7308	3,4%	121	966	11,1%	47	2216	2,1%
MPPMU	270	2020	11,8%	64	879	6,8%	6	506	1,2%
PRAGUE	791	442	64,2%	12	161	6,9%	11	1502	0,7%
UNI-DORTMUND	257	169	60,3%	281	114	71,1%	106	100	51,5%
UNI-FREIBURG	1303	2097	38,3%	64	840	7,1%	29	1941	1,5%
WUPPERTALPROD	996	1446	40,8%	6	101	5,6%	4	36	10,0%

Should be aiming to less then 1% failure rate !

Recent HC test - Eventrate

	Panda		dcap/rfio		FileStager	
	Mean (Hz)	Error (Hz)	Mean (Hz)	Error (Hz)	Mean (Hz)	Error (Hz)
CSCS	4,1	4,1	4,1	1,7	9,9	7,9
CYF	5,8	4,4	0,9	0,6	0,0	0,0
DESY-HH	6,3	4,1	4,1	2,6	17,2	6,9
DESY-ZN	8,7	5,0	12,0	4,0	14,1	6,6
FZK	8,2	5,3	4,4	3,0	14,0	7,5
GOEGRID	8,5	6,1	7,5	3,7	13,1	7,8
HEPHY-UIBK	4,2	2,9	2,2	2,7	17,4	7,3
LRZ-LMU	9,6	5,3	12,3	5,2	16,9	7,9
MPPMU	5,9	3,8	11,1	33,0	4,8	4,3
PRAGUE	5,0	2,7	1,6	1,4	6,7	4,9
UNI-DORTMUND	4,5	4,2	0,0	0,0	0,3	1,5
UNI-FREIBURG	5,5	5,1	3,9	2,1	13,8	7,6
WUPPERTALPROD	3,9	3,5	2,0	0,8	3,8	4,6

The more the better !

Závěry

- Výsledky z pohledu uživatele nejsou ideální
  - ale byly by mnohem lepší v rámci přislíbených zdrojů
- Plánován upgrade páteřní sítě na 10 Gbps
- Mnoho dalších uživatelských úloh během testů
  - stále nemáme nastaveny poměry mezi různými typy úloh
- Lokální zdroje jsou pro analýzu velmi vhodné