Stakeholder priority on the Shared Farm

10-26-2021: As we migrating to SDF, decommissioning old hardware and RHEL6, we will no longer actively update the fair shares in this page - some of the major stakeholders no longer use LSF batch system in large scale.

The Shared (General) Farm consists of several physical clusters that are available to all SLAC users. The cluster hardware was purchased incrementally over several years by various stakeholder groups. Each physical cluster is based on a specific hardware model but all hosts in the farm run 64bit RHEL /CentOS6. The LSF "general" queues feed user's jobs to the shared farm. The stakeholders have associated LSF user groups with a fairshare (scheduling priority) that reflects their cluster investment. This ensures that stakeholders will always get some runtime on the cluster when utilization is high. Users that are not members of stakeholder groups still have the ability to run jobs "for free". There is a superset fairshare group "AllUsers" that includes all SLAC users. A non-stakeholder must compete for priority with all other users running jobs on the shared farm. This may be acceptable for some, but production environments may demand priority scheduling. The free "AllUsers" fairshare is actually subsided by the paying stakeholders. A stakeholder's fairshare value is derived from the compute power they have purchased. A HS06 CPU benchmark is calculated for each cluster server model.

Stakeholder Investments

Cluster	# of hosts	# cores/host	# of cores	# Batch slots	HS06/slot	HS06	Owner/Group	Purchased date	Notes
fell	-	-	-	-	-	-	-	2007	IGNORE - obsolete hardware in run-to-fail mode
~~hequ~~	~~39~~	~~8~~	~~312~~	~~312~~	~~15.01~~	~~4683.12~~	~~ATLAS~~	~~9/23/2009~~	RUN-to-FAIL. RHEL6
~~hequ~~	~~76~~	~~8~~	~~608~~	~~608~~	~~15.01~~	~~9126.08~~	~~Fermi~~	~~9/23/2009~~	RUN-to-FAIL. RHEL6
~~hequ~~	~~77~~	~~8~~	~~616~~	~~616~~	~~15.01~~	~~9246.16~~	~~Babar~~	~~9/23/2009~~	RUN-to-FAIL. RHEL6
~~dole~~	~~38~~	~~12~~	~~456~~	~~456~~	~~13.77~~	~~6279.12~~	~~ATLAS~~	~~10/19/2010~~	RUN-to-FAIL. RHEL6
kiso(i)	68	24	1632	1360	9.97	13559.2	ATLAS	9/23/2011	RUN-to-FAIL. CentOS7 HT enabled/ 20 slots per host
bullet	77.5	16	1240	1240	16.05	19902	PPA	12/5/2012	RUN-to-FAIL. RHEL6 for MPI use - do not map to fairshare
bullet	76.25	16	1220	1220	16.05	19581	Fermi	12/5/2012	RUN-to-FAIL. RHEL6
bullet	17.75	16	284	284	16.05	4558.2	Geant	12/5/2012	RUN-to-FAIL. RHEL6
bullet	47.25	16	756	756	16.05	12133.8	ATLAS	2/13/2014	RUN-to-FAIL. RHEL6
bullet	49.25	16	788	788	16.05	12647.4	PPA	2/13/2014	RUN-to-FAIL. RHEL6
bullet	19.5	16	312	312	16.05	5007.6	Theory	2/13/2014	RUN-to-FAIL. RHEL6
bullet	32.5	16	520	520	16.05	8346	Beamphysics	10/2/2014	RUN-to-FAIL. RHEL6. for MPI use - do not map to fairshare
deft(i)	22	24	528	528	14.97	7904	ATLAS	1/1/2016	CentOS7 HT disabled, two purchases at 11/2015 and 1/2016
deft(i)	7	24	168	168	14.97	2514.96	Fermi	6/1/2016	CentOS7
bubble(i)	8	36	288	288	21.64	6232	Fermi	3/2018	CentOS7. HT disabled
bubble(i)	12	36	432	432	21.64	9348.48	Beamphysics	9/2018	CentOS7.Priority for Beamphysics MPI use - do not map to fairshare

- i) CentOS 7 with Singularity

bullet MPI users

The bulletmpi queues are not associated with these fairshare policies. MPI parallel jobs reserve cores across multiple hosts. This does not work well with fairshares because the fairshare formula takes reserved cores into consideration. Large parallel jobs would never start because a job's dynamic scheduling priority would drop before it reserves all the cores it needs to get started. The bulletmpi queues have a higher queue-level priority than all of the fairshare general queues. For this reason we do not assign general queue fairshares for MPI investments.

AllUsers 'Tax'

The stakeholders subside the AllUsers ("free") fairshare by paying a 15% tax on their shares:

Group	HS06	HS06 (after tax)
ATLAS	44559	37875
Fermi	31222	26539
Babar	9246	7859
Geant	4558	3874
Theory	5008	4257

PPA others	12647	10750

The 15% tax results in 16086 shares for AllUsers

LSF fairshare user groups

The stakeholders are responsible for distributing their shares among their respective LSF groups. The table below should match the production LSF config:

USER/GROUP	HS06_SHARES	%HS06_SHARES	HS06_OWNER
atlasgrp	37875	35.32%	ATLAS
babarAll	7859	7.33%	BaBar
glastdata	854	0.80%	Fermi
glastusers	25320	23.61%	Fermi
glastgrp	366	0.34%	Fermi
geantgrp	3874	3.61%	Geant
luxlz	3500	3.26%	PPA others
cdmsdata	2000	1.86%	PPA others
lcdprodgrp	1100	1.03%	PPA others
exoprodgrp	1500	1.40%	PPA others
hpsprodgrp	1000	0.93%	PPA others
rpgrp	500	0.47%	PPA others
lcd	600	0.56%	PPA others
exousergrp	550	0.51%	PPA others
rdgrp	0	0.00%	PPA others
theorygrp	4257	3.97%	Theory
All Users	16086	15.00%	Everyone (15% tax)

Decommissioning clusters

Fairshares should be subtracted from stakeholder groups and the associated AllUsers tax when hardware is declared obsolete and put into run-to-fail mode. The fairshare distribution should always reflect stakeholder investment in current production cluster resources.

Chargeback model for sustainable hardware lifecycle

This effort paves the way for a possible chargeback model where Computing Division could lease cluster hardware and charge stakeholders a rate for fairshares.