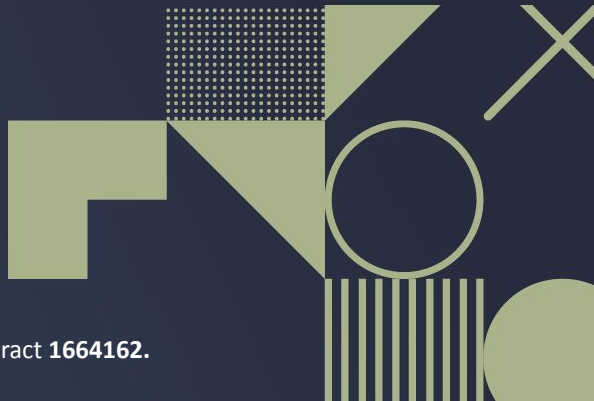
A Lightweight GPU Monitoring Extension for Pegasus Kickstart

George Papadimitriou, Ewa Deelman

Workshop on Workflows in Support of Large-Scale Science (WORKS)
November 15th, 2021

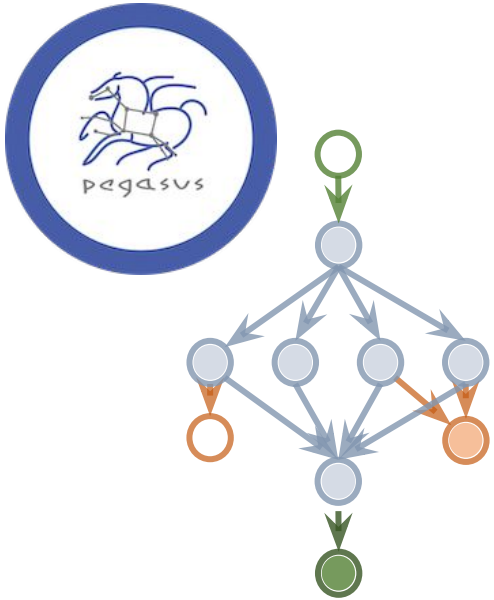


This work was funded by the US Department of Energy under contract **DE-SC0012636M** and by NSF under contract **1664162.**

Pegasus Kickstart: What is it?

Pegasus Kickstart (or Kickstast for short)

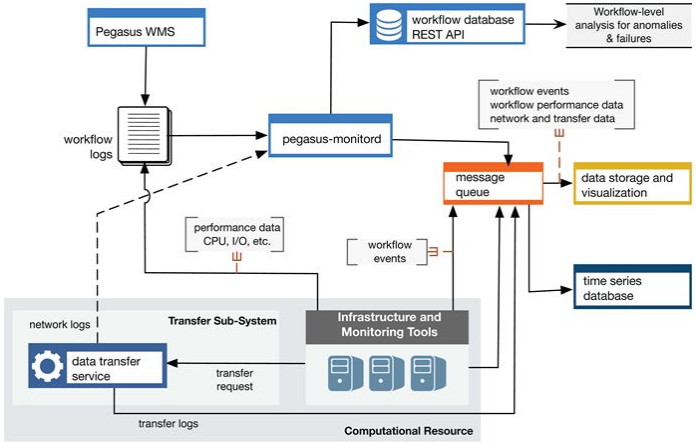
- A lightweight job wrapper
- Implemented in C
- Captures performance and provenance data
- YAML output
- Part of the Pegasus workflow management system tools



Kickstart in Panorama: What's new?

Pegasus Panorama extends Kickstart

- Online collection of performance statistics (traces) as captured by Linux's procfs
- Correlates them with workflow jobs
- Interval as fast as 1 second
- AMQP publish capability
- JSON output



panorama360

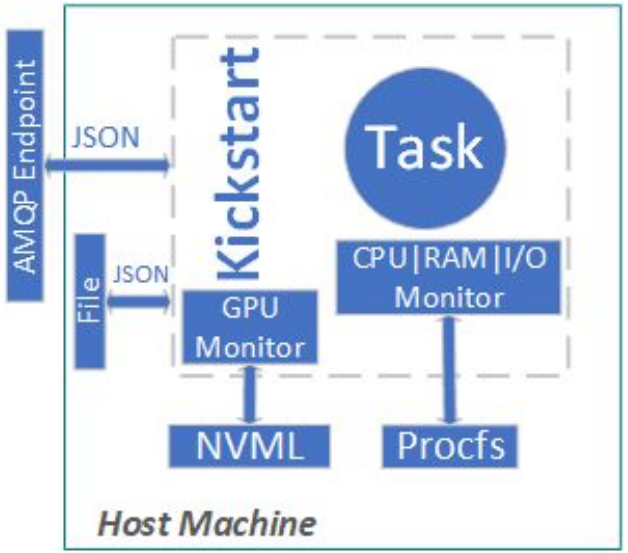
https://panorama360.github.io

Now with Nvidia Monitoring Support !

Kickstart: Nvidia Monitoring - Approach

- Leverages Nvidia's Monitoring Library (NVML)
  - C-based API
  - Interfacing with the Nvidia driver
- Kickstart implements a lightweight C wrapper for NVML
- Kickstart polls for new GPU statistics as fast as 1 second
- Multi-threading support

https://docs.nvidia.com/deploy/nvml-api/index.html



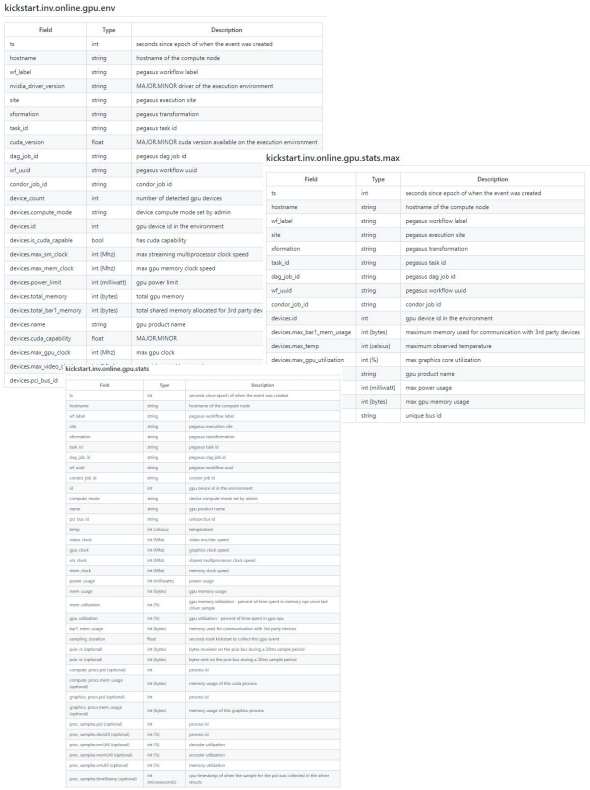
Kickstart online monitoring

Kickstart: Nvidia Monitoring - Events

- **kickstart.inv.online.gpu.env:** Contains information about the GPU environment - produced once at the beginning of the job

- **kickstart.inv.online.gpu.stats:** Contains a snapshot of the GPU counters - produced throughout the execution

- **kickstart.inv.online.gpu.stats.max:** Summary of max values observed - produced after the job finishes



kickstart.inv.online.gpu.env

Field	Type	Description
ts	int	seconds since epoch of when the event was created
hostname	string	hostname of the compute node
wf_label	string	pegasus workflow label
nvidia_driver_version	string	MAJOR.MINOR driver of the execution environment
site	string	pegasus execution site
xformation	string	pegasus transformation
task_id	string	pegasus task id
cuda_version	float	MAJOR.MINOR cuda version available on the execution environment
dag_job_id	string	pegasus dag job id
wf_uuid	string	pegasus workflow uuid
condor_job_id	string	condor job id
device_count	int	number of detected gpu devices
devices.compute_mode	string	device compute mode set by admin
devices.id	int	gpu device id in the environment
devices.is_cuda_capable	bool	has cuda capability
devices.max_sm_clock	int (Mhz)	max streaming multiprocessor clock speed
devices.max_mem_clock	int (Mhz)	max gpu memory clock speed
devices.power_limit	int (milliwatt)	gpu power limit
devices.total_memory	int (bytes)	total gpu memory
devices.total_bar1_memory	int (bytes)	total shared memory allocated for 3rd party dev
devices.name	string	gpu product name
devices.cuda_capability	float	MAJOR.MINOR
devices.max_gpu_clock	int (Mhz)	max gpu clock
devices.max_video_c		
devices.pci_bus_id		

kickstart.inv.online.gpu.stats.max

Field	Type	Description
ts	int	seconds since epoch of when the event was created
hostname	string	hostname of the compute node
wf_label	string	pegasus workflow label
site	string	pegasus execution site
xformation	string	pegasus transformation
task_id	string	pegasus task id
dag_job_id	string	pegasus dag job id
wf_uuid	string	pegasus workflow uuid
condor_job_id	string	condor job id
devices.id	int	gpu device id in the environment
devices.max_bar1_mem_usage	int (bytes)	maximum memory used for communication with 3rd party devices
devices.max_temp	int (celsius)	maximum observed temperature
devices.max_gpu_utilization	int (%)	max graphics core utilization
	string	gpu product name
	int (milliwatt)	max power usage
	int (bytes)	max gpu memory usage
	string	unique bus id

kickstart.inv.online.gpu.stats

Field	Type	Description
ts	int	seconds since epoch of when the event was created
hostname	string	hostname of the compute node
wf_label	string	pegasus workflow label
site	string	pegasus execution site
xformation	string	pegasus transformation
task_id	string	pegasus task id
dag_job_id	string	pegasus dag job id
wf_uuid	string	pegasus workflow uuid
condor_job_id	string	condor job id
id	int	gpu device id in the environment
compute_mode	string	device compute mode set by admin
name	string	gpu product name
pci_bus_id	string	unique bus id
temp	int (celsius)	temperature
video_clock	int (Mhz)	video encoder speed
gpu_clock	int (Mhz)	graphics clock speed
sm_clock	int (Mhz)	shared multiprocessor clock speed
mem_clock	int (Mhz)	memory clock speed
power_usage	int (milliwatts)	power usage
mem_usage	int (bytes)	gpu memory usage
mem_utilization	int (%)	gpu memory utilization - percent of time spent in memory ops since last nvsmi sample
gpu_utilization	int (%)	gpu utilization - percent of time spent in gpu ops
bar1_mem_usage	int (bytes)	memory used for communication with 3rd party devices
sampling_duration	float	seconds task kickstart to collect this gpu event
pcie_rx (optional)	int (bytes)	bytes received on the pcie bus during a 20ms sample period
pcie_tx (optional)	int (bytes)	bytes sent on the pcie bus during a 20ms sample period
compute_procs.pid (optional)	int	process id
compute_procs.mem_usage (optional)	int (bytes)	memory usage of this cuda process
graphics_procs.pid (optional)	int	process id
graphics_procs.mem_usage (optional)	int (bytes)	memory usage of this graphics process
proc_samples.pid (optional)	int	process id
proc_samples.sm2dt (optional)	int (%)	process id
proc_samples.enc2dt (optional)	int (%)	encoder utilization
proc_samples.mem2dt (optional)	int (%)	encoder utilization
proc_samples.cmd2dt (optional)	int (%)	memory utilization
proc_samples.timeStamp (optional)	int (microseconds)	cpu timestamp of when the sample for the pid was collected in the driver struct

Kickstart: Nvidia Monitoring - Levels

- **DEFAULT:** Reports general status of the GPU and statistics of the CUDA kernels
- **MONITORING_PCIE:** Enables monitoring of the pcie bus for data transmission to and from the GPU
- **GRAPHICS_PROCS:** Captures statistics of the graphics (video) processes

Monitoring Level	Measurement Time
DEFAULT	~6ms
MONITORING_PCIE	~228ms

More information on GitHub

https://github.com/pegasus-isi/pegasus/tree/panorama/src/tools/pegasus-kickstart/nvidia

Kickstart: Nvidia Monitoring - Invocation

- **KICKSTART_MON_URL:** Sets the location the statistics will be saved (AMQP or file)
- **-m INT:** Enables online monitoring with interval INT
- **-G:** Enables default gpu monitoring
- If you want to use with Pegasus consider referencing
  - https://pegasus.isi.edu/documentation/manpages/pegasus-kickstart.html?highlight=kickstart.
  - https://pegasus.isi.edu/presentations/2019/pegasus-office-hours-monitoring.pdf

```
export KICKSTART_MON_URL = \
        rabbitmq://[USERNAME:PASSWORD]@hostname[:port]/api/exchanges/[VIRTUAL_HOST]/[EXCHANGE_NAME]/publish
or
export KICKSTART_MON_URL = file://filename
pegasus-kickstart <args> -G -m 10 ./exec
```

Listing 1: Example invocation of GPU monitoring

FAQ: Do I Need to Use The Entire Pegasus WMS?

- No!
- **Kickstart is standalone and can be invoked independently!**
  - But if you are using Pegasus you automatically benefit from it ;)
- **Prebuilt versions of the GPU enabled Kickstart available online**
- **Offered through the lightweight pegasus worker package**

http://download.pegasus.isi.edu/pegasus/5.1.0panorama

QUESTIONS ?

George Papadimitriou
University of Southern California
georgpap@isi.edu

https://pegasus.isi.edu

https://panorama360.github.io

https://github.com/pegasus-isi/pegasus/tree/panorama

This work was funded by the
US Department of Energy
under contract **DE-SC0012636M,**
and by NSF under contract **1664162.**