Introduction to Slurm

The **Slurm Workload Manager** (formerly known as *Simple Linux Utility for Resource Management or SLURM*), is a free and open-source job scheduler for Linux and Unix-like kernels, used by many of the world's supercomputers and computer clusters. It provides three key functions.

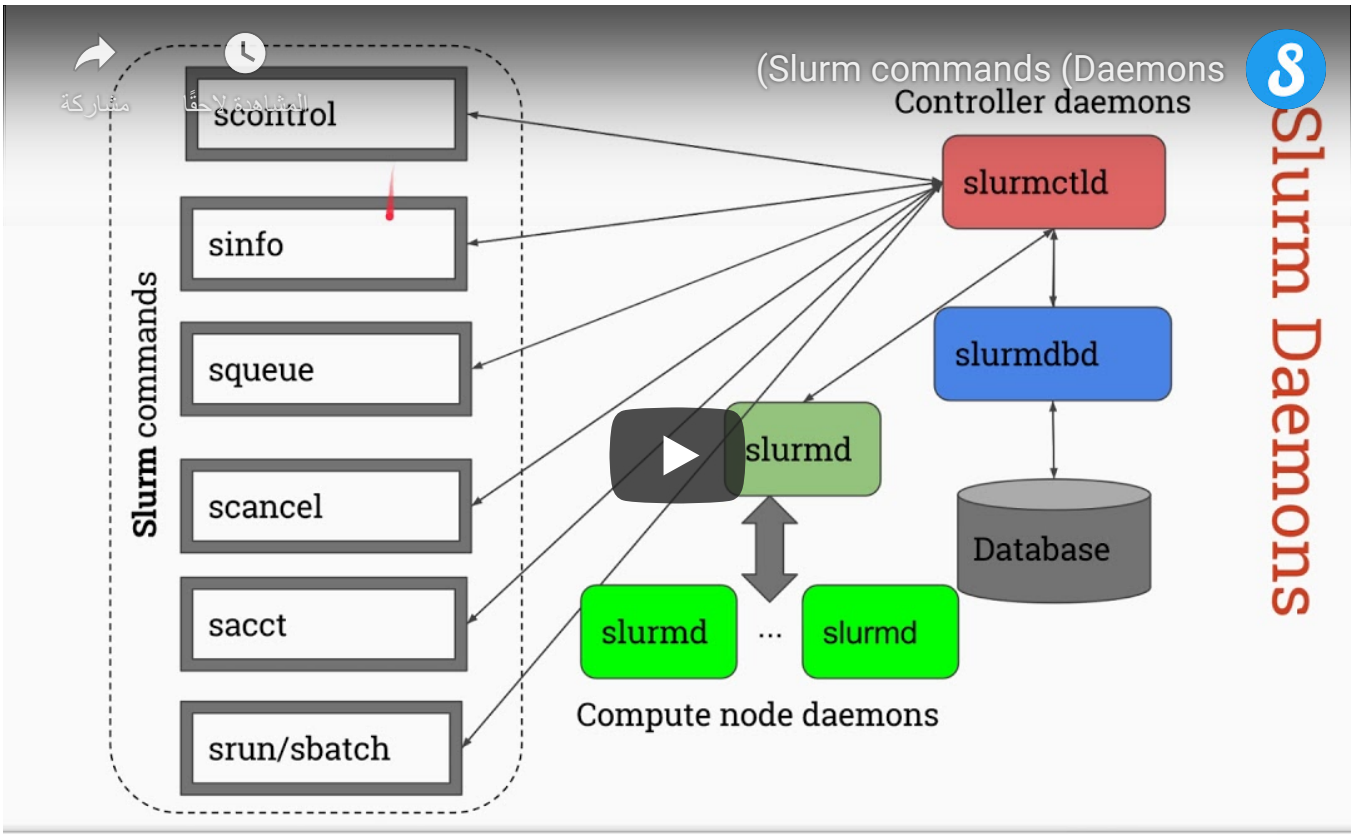
- First, it **allocates exclusive and/or non-exclusive access** to resources (computer nodes) to users for some duration of time so they can perform work.
- Second, it provides a **framework for starting, executing, and monitoring** work (typically a parallel job such as `MPI`) on a set of allocated nodes.
- Finally, it arbitrates **contention for resources** by managing a queue of pending jobs.

Slurm is the workload manager on about 60% of the TOP500 supercomputers, including Tianhe-2 that, until 2016, was the world's fastest computer.

History

Slurm began development as a collaborative effort primarily by Lawrence Livermore National Laboratory, **SchedMD**, Linux NetworX, Hewlett-Packard, and Groupe Bull as a Free Software resource manager in the 2010s. It was inspired by the closed source Quadrics RMS and shares a similar syntax. The name is a reference to the soda in Futurama!

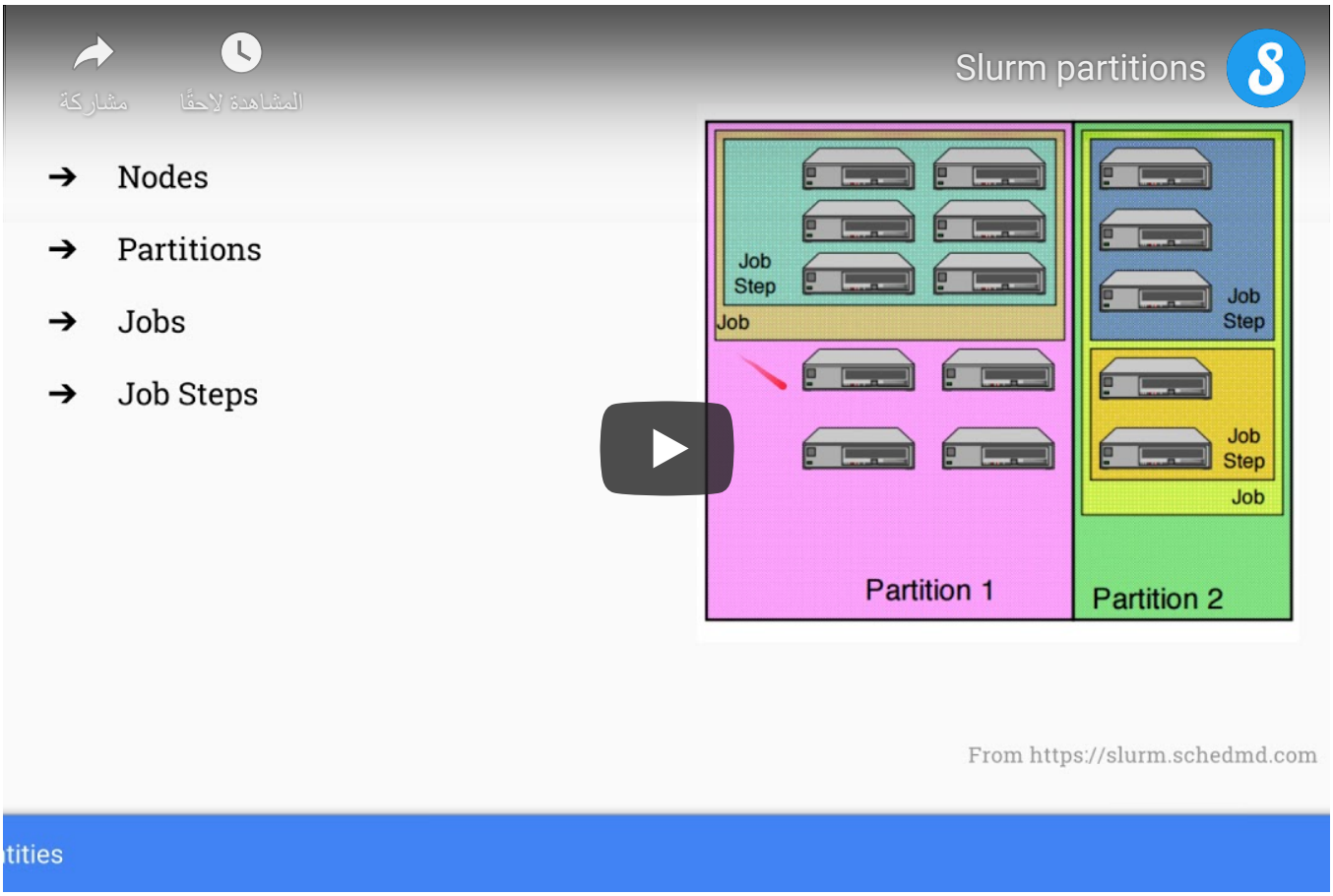
Components of Slurm workload manager



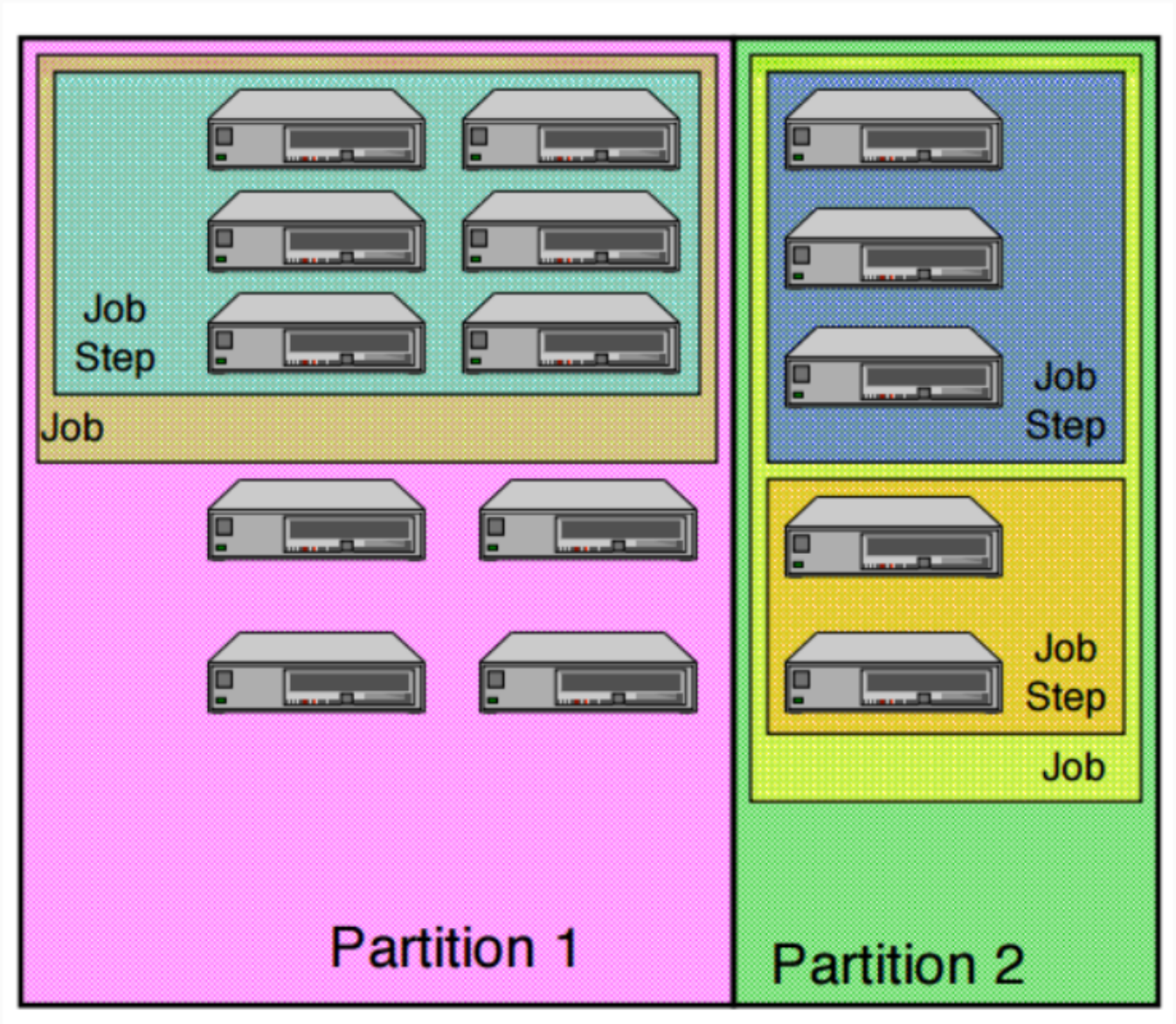
Reproduced from
https://slurm.schedmd.com

Slurm is a sophisticated batch scheduler capable of satisfying the requirements of many large computer centers. Slurm consists of a `slurmd` daemon running on each compute node and a central `slurmctld` daemon running on a management/ master node. The `slurmd` daemons provide fault-tolerant hierarchical communications. The user commands include: `sacct`, `salloc`, `sattach`, `sbatch`, `sbcast`, `scancel`, `scontrol`, `sinfo`, `smap`, `squeue`, `srun`, `strigger` and `sview`. All of the commands can run anywhere in the cluster.

Slurm entities



In contrast to **PBS**, **Slurm** in revolves around four entities: `nodes`, `partitons` (similar to `queues` in the PBS, but not the same!), `jobs` and `job steps`.



Slurm entities distributed of the compute resources

Where `nodes` are the compute resource in Slurm, `partitions` are the group of nodes into logical (possibly overlapping) sets, `jobs` are the allocations of resources assigned to a user for a specified amount of time, and `job steps` are the sets of (possibly parallel) tasks within a job.

Note that the Slurm `partitions` can be considered job `queues`, each of which has an assortment of constraints such as job `size limit`, job `time limit`, `users` permitted to use it, etc.