PN-08

Wet GPU Software Stack

Denis Lalaj, Harjot Grewal, Jacob Yang, Jerry Shao, Phi Lam

The Client

samnian

Emerging technologies **software company** based in Vancouver. In-house product development for **AI applications** and **cleantech**.

The Problem

- Computing centers (HPCs) handle **heavy workloads**
- Heavy workloads can lead to **long queue** times for users
- Upgrading computing infrastructure can be **infeasible and costly**

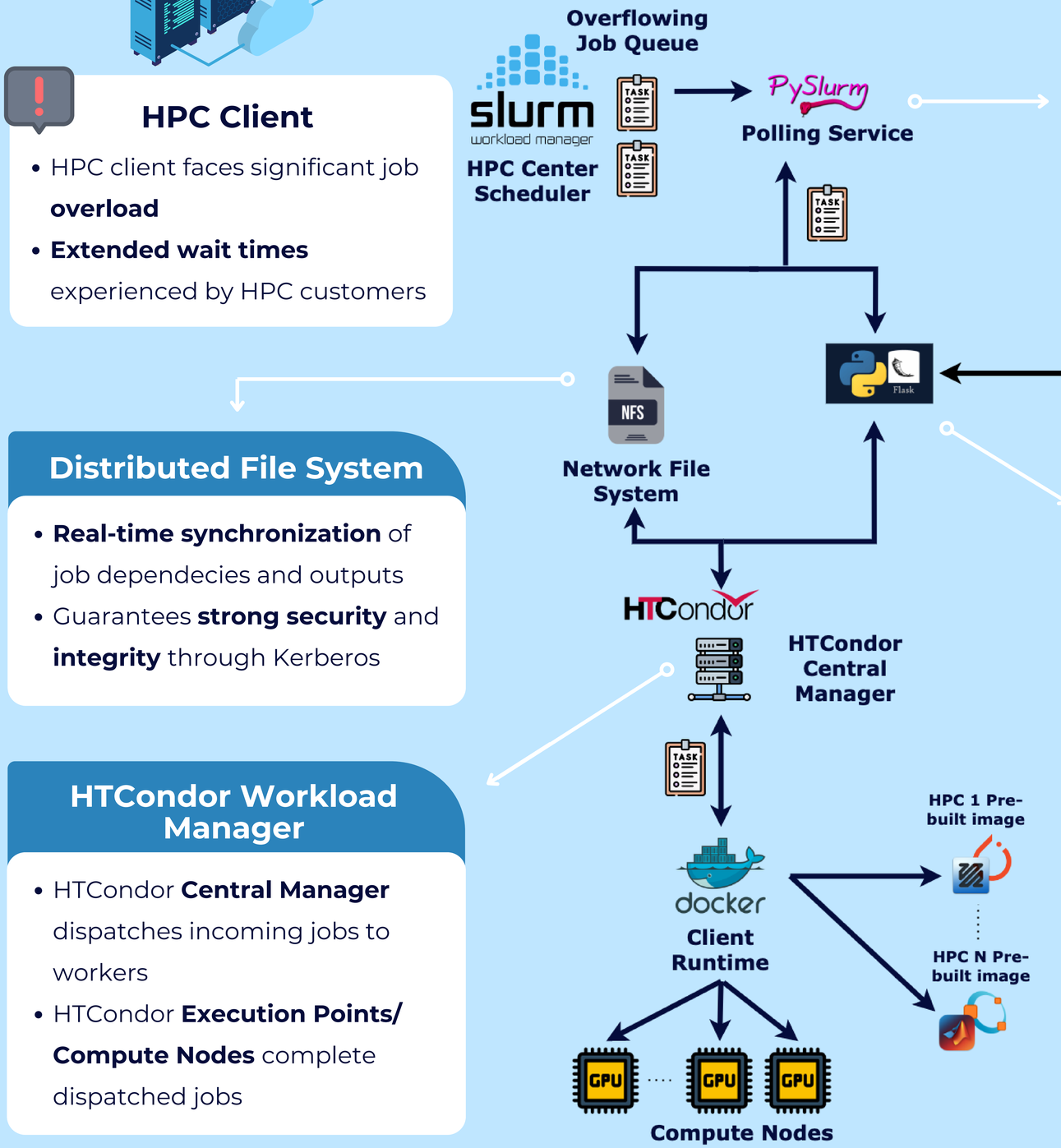
The Solution

- Outsourcing service as **support infrastructure** for HPC centers
- Reroutes **overflowing jobs** from swamped HPC centers to our cluster
- Equipped with **cleantech hardware** to align with environmental goals

HPC Client

- HPC client faces significant job **overload**
- **Extended wait times** experienced by HPC customers

System Architecture



Distributed File System

- **Real-time synchronization** of job dependecies and outputs
- Guarantees **strong security** and **integrity** through Kerberos

HTCondor Workload Manager

- HTCondor **Central Manager** dispatches incoming jobs to workers
- HTCondor **Execution Points/ Compute Nodes** complete dispatched jobs

HPC Monitoring Service

- PySlurm daemon in HPC **detects jobs** from **overflowing queue**
- **Dequeues and forwards** job details to our system over **NFS**

RESTful Web Server

- Provides **portable interface** for **job submission and monitoring**
- **High granularity** tracking of **worker node health** to time series database

Containerized Runtime

- Docker containers reproduce HPC **execution environment** using pre-built images
- Containers provide **isolation** from the host operating system