DARE-BigNGS : A Science Gateway Model for Enabling Scalable Next-generation Sequencing (NGS) Data Analytics

Joohyun Kim

Center for Computation and Technology

Louisiana State University

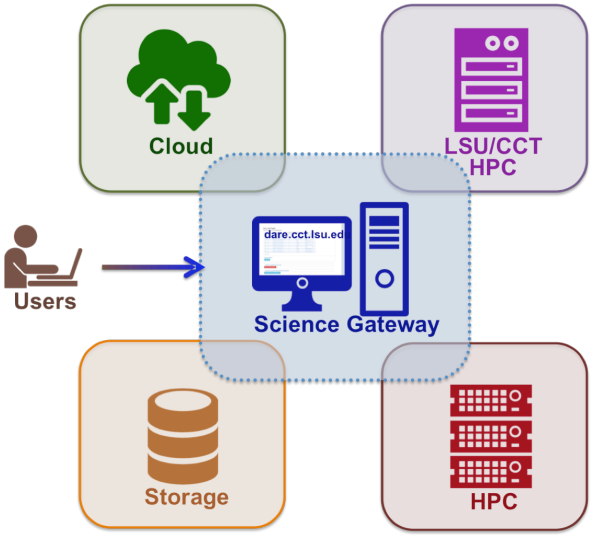
Background

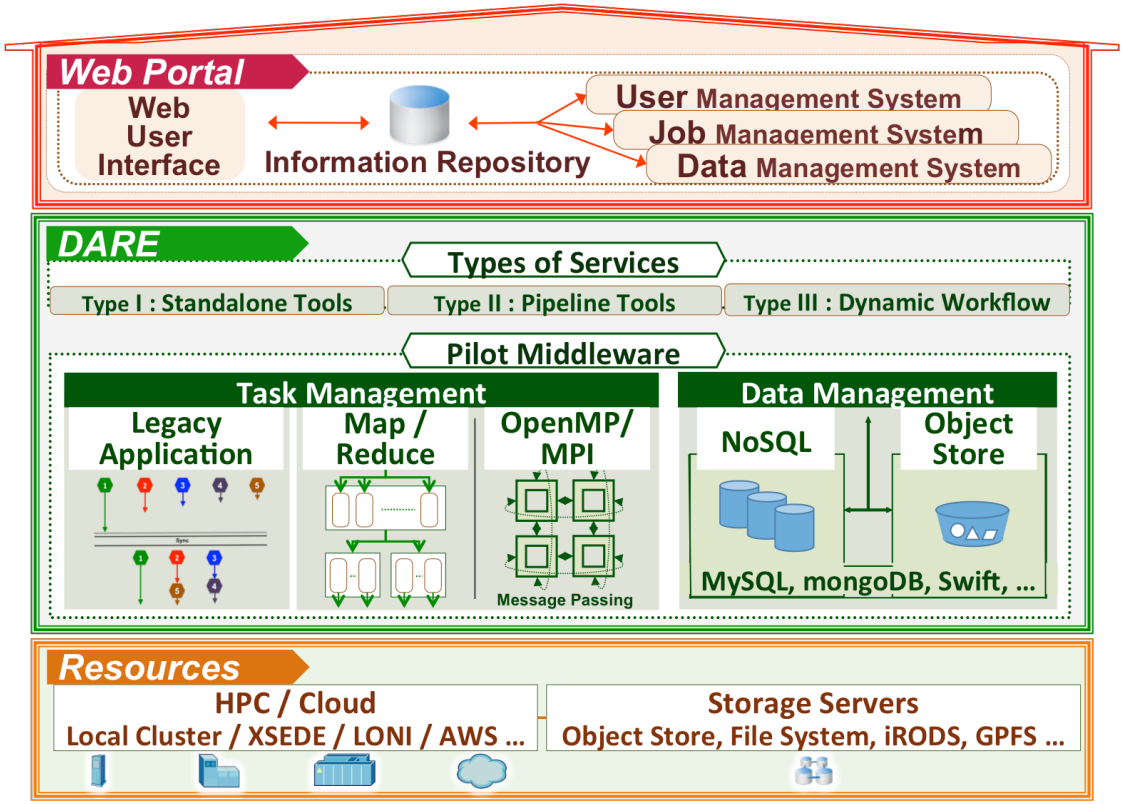
- Science Gateway is an effective means for individual researchers or small groups to utilize computing resources that they do not own
- Lightweight modular three layer architecture comprising a web portal, DARE-based middleware, and distributed resources including HPC and Cloud environments
- Distributed Application Runtime Environment (DARE) and developed components with open source tools
- Next-generation Sequencing (NGS) data anayltics and downstream analyses as specifically targeted services

Contents

- INTRODUCTION
  - DARE-NGS, built upon a model for gateway projects aiming to serve scalable data analytics
- RECENT DEVELOPMENTAL ACHIEVEMENTS
  - Front end : recent advances using web application frameworks and tools for rich user experiences
  - Back end : support of Hadoop applications, on-demand cloud environments, docker container technology, and other additions
- PILOT PROJECT WITH IBM
  - Special use case scenarios with Delta, IBM Power8 system with IBM Spectrum Software and IBM Bluemix cloud

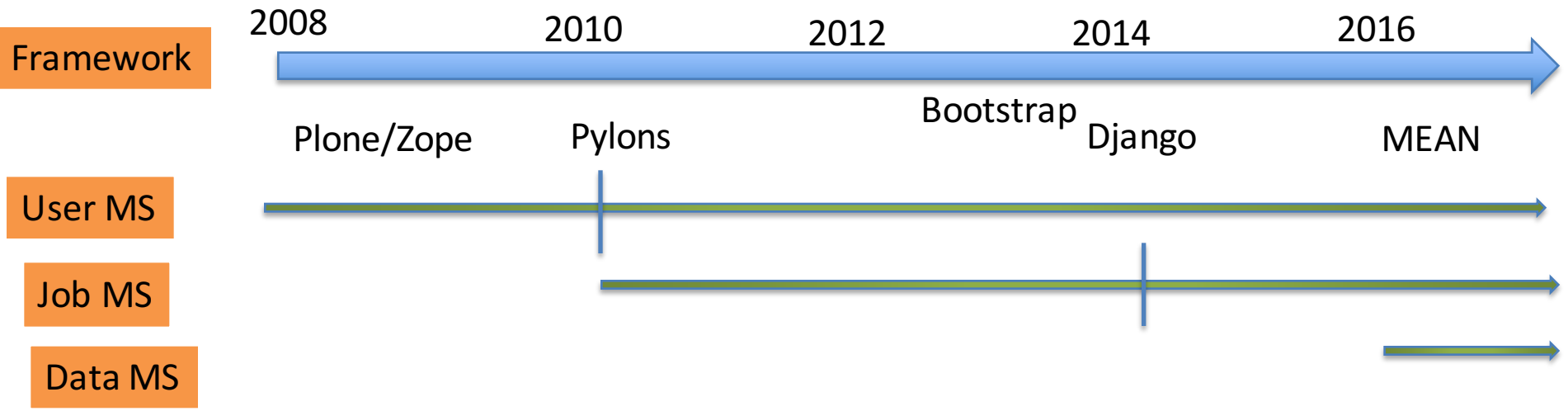
Model : Overall Architecture





Model : Brief History

Front End : Web application



Back end : Distributed resources and distributed application framework

LOVE PURPLE
LIVE GOLD

Overall Design Objectives

- **Web portal** : Lightweight, modular, extensible, and provision of rich user experience
- **DARE** : **Scalability** across HPCs (Local and XSEDE) and Clouds (Amazon EC2, IBM Bluemix, and OpenStack systems)

DARE-BigNGS

- **WHY?**

  NGS data analytics require a scalable solution

  - **Ever-growing data sets (Volume) and complexity of analytics with noisy data sets (Veracity and Variety)**

- **WHAT?**

  Provision of services of NGS data analytics for end-to-end solution

- **HOW?**

  Distributed Runtime Environment framework for massive data sets and massive task management

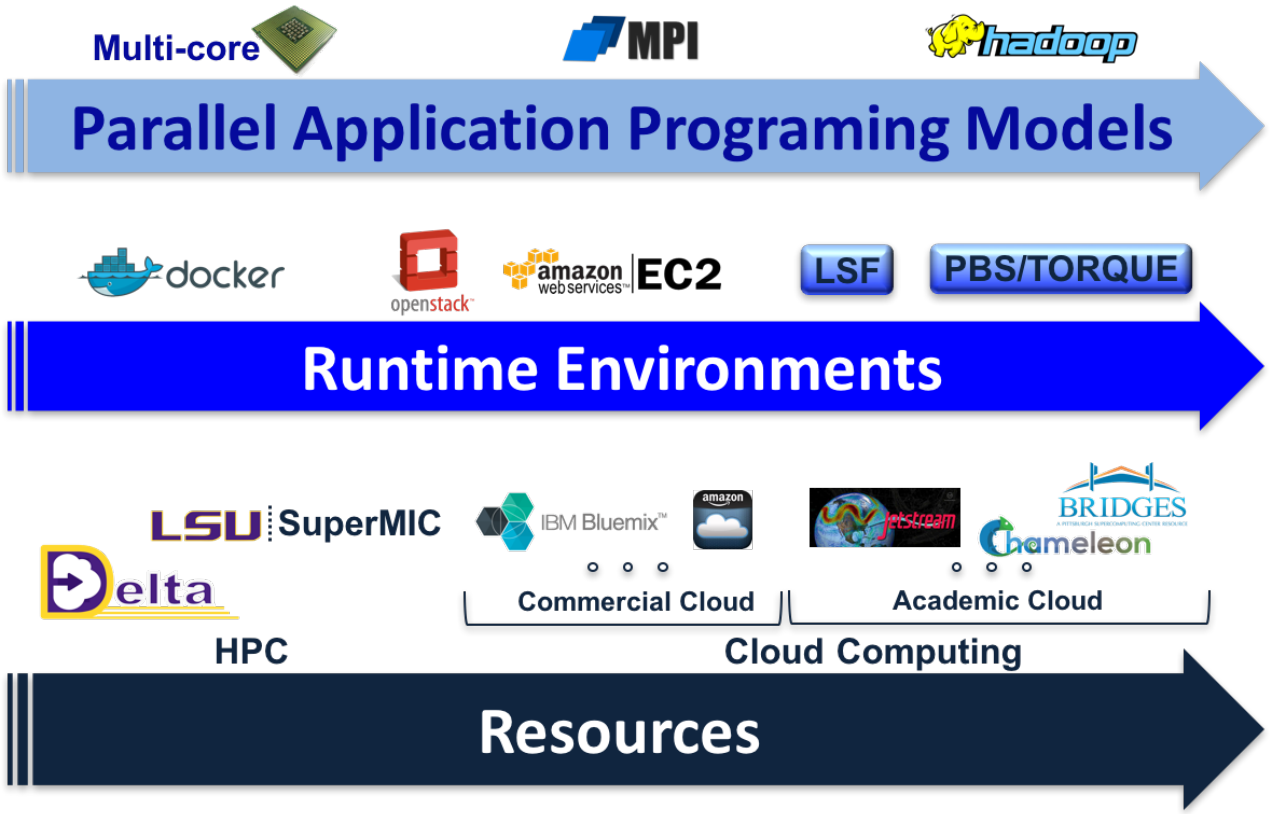
One More

In fact, reminiscent themes of data science topics!

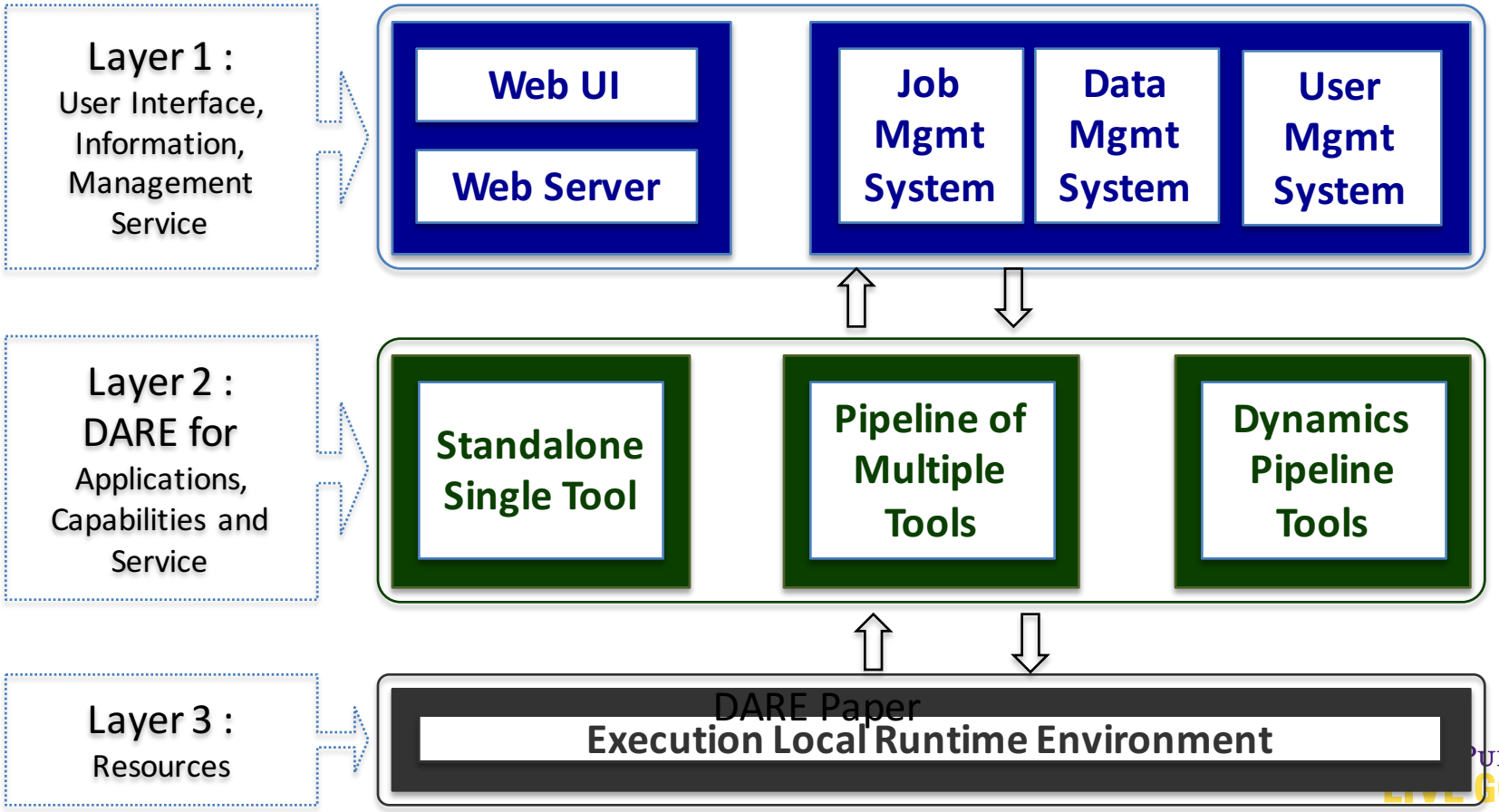
- Large data sets
- Massive tasks for model complexity and noisy data
  - Many machine learning methods need to search optimal hyper-parameters
  - Ensemble approaches (Random Forest, Boosting, and others) are generally better
- Solution by leveraging
  - Distributed heterogeneous environment
  - Distributed Big Data models such as MapReduce, Spark, and others

DARE over HPCs and Clouds

Uniform interface to support distributed heterogeneous HPC and Cloud resources and their local environments

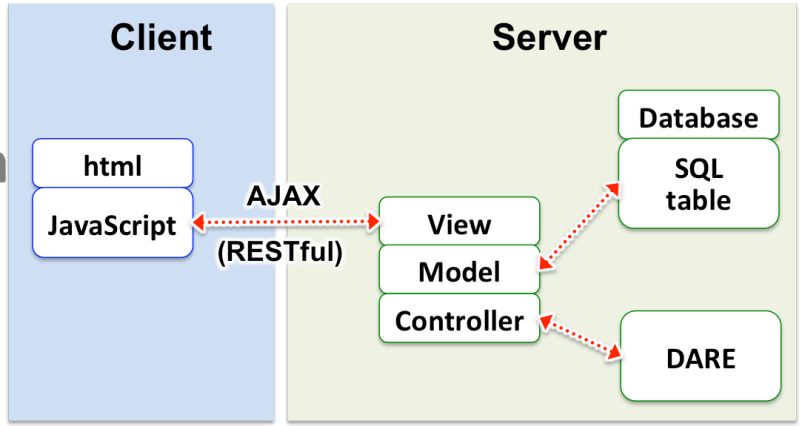


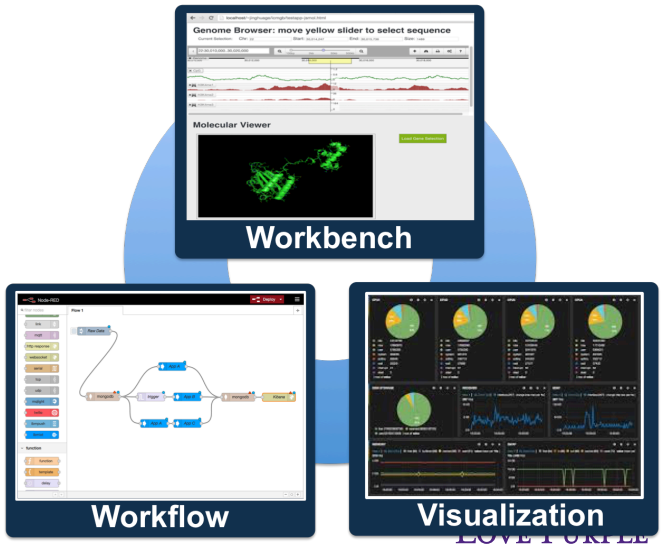
DARE-BigNGS Structure



Web Portal

- Main Web Application (Authentication and authorization)
- Web Client
  - Bootstrap as a design framework
  - JavaScript frameworks (jQuery, AnglularJS)
  - AJAX/REST client (AngularJS)
  - WebSockets
- Web Server
  - MVC paradigm with a web application framework
  - Python web application framework (Pylons, Django, and Zope/Plone)
  - JavaScript framework (MEAN: express/node)
  - REST/WebSocket Server side . [Workbench Design]





LIVE GOLD

Service Management System

- Job Management System, Data Management System, Resource Management System
- DARE for distributed job and data management
- Database back-ends : MySQL and MongoDB
- Object store : SWIFT
- UI Workbench for Job submission, monitoring, visualization of information
- Pilot framework for workflow
- Local runtime environments of remote systems
  - Legacy HPC environments
  - Virtualization clouds (EC2, IBM Bluemix, Chameleon, JetStream)
  - Docker container

DARE supports scalable multi-omics data analytics

Multi-faceted genomic/epigenomic landscape or networks of living cells :

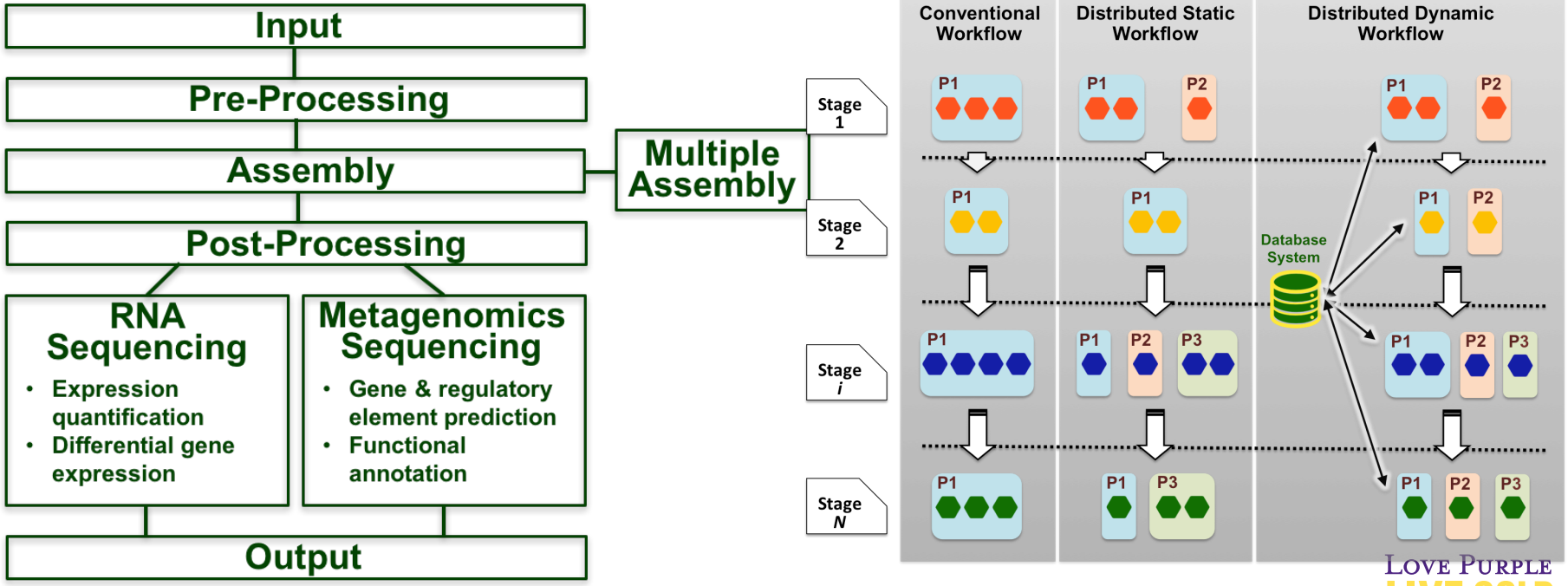
- Multiple samples (cohort size)
- Multi-platform protocols : WGS, RNA-seq, Methyl-seq, ChIP-seq, DNase-seq, Hi-C, miRNA-seq, and more.
- Common challenges across a diverse set of omics approaches as well as the holy grail challenge (integrative solution)

DARE supports scalable multi-omics data analytics

- Service I. Transcriptom/metagenome analysis (bigTrans/bigMeta)
- Service II. RNA-Seq Pipeline for Differential Gene Expression analysis
- Service III. Somatic Mutation Discovery (SMD) pipeline
- Service IV. DNase-seq data analysis using Deep Learning

DARE means optimized execution of distributed applications

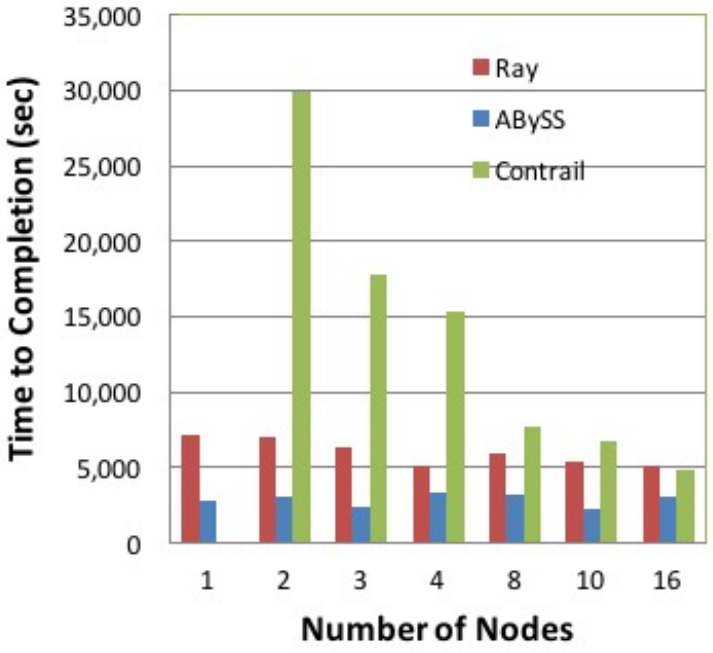
Service for Transcriptome/metagenome Profiling : :bigTrans/bigMeta

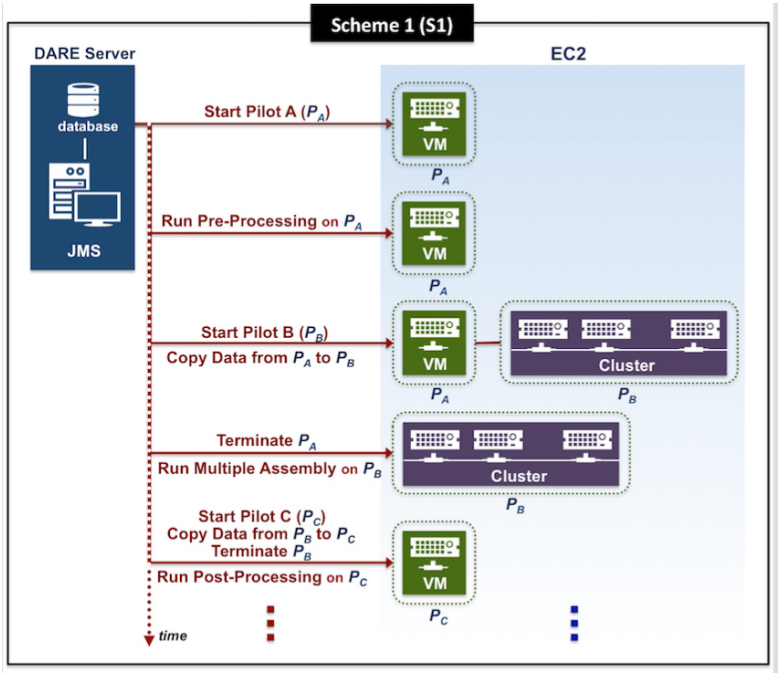


LOVE PURPLE
LIVE GOLD

DARE means optimized execution of distributed applications

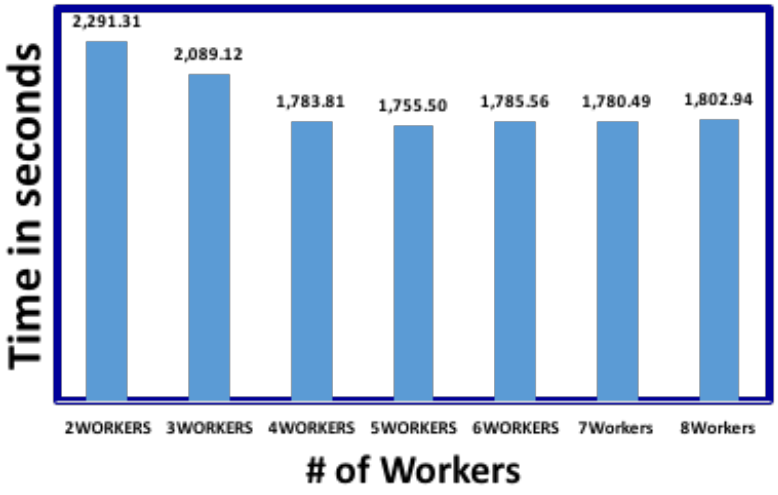
- Pilot-based workflow execution over virtualization clouds (EC2, IBM Bluemix, and OpenStack (Chameleon

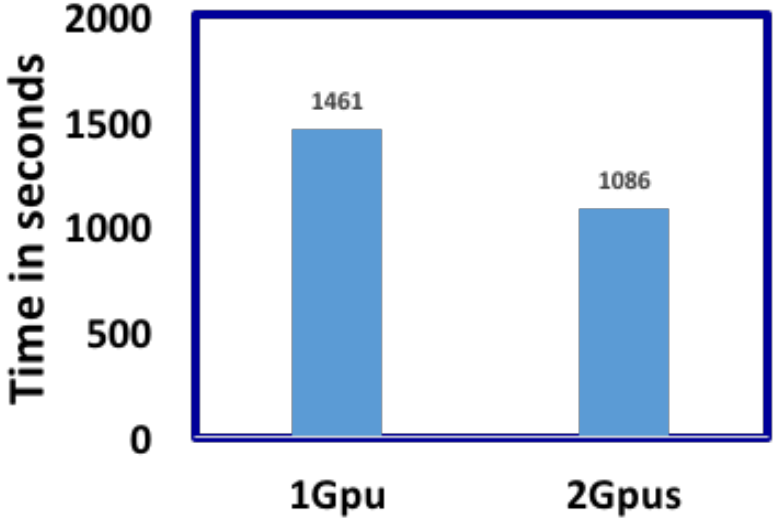




DARE means optimized execution of distributed applications

- Support of BigData applications (Hadoop models) and Deep Learning





(Tensoflow (DeepLearning) Implemented DNase-seq)

Job Execution with Docker

- Docker container image packaged and configured for a target analytics application
- Support of MPI applications with multiple Docker containers
- Support of Hadoop/spark applications
- SWIFT object store system for Docker image repository
- Rapid prototyping of a service and efficient packaging
- Two options for DARE to access a container
  - Option A : RESTful API-based pilot agent
  - Option B : Legacy pilot-based agent with ssh

Job Execution with Docker

Option A details

- Node-RED can be a pilot master
- Docker containers as pilot agents
- Docker image a the target application and REST web server and APIs
- MongoDB is the centralized back-end and use Docke Compose to up the pilot instance (mutliple container docker instance)
- (Pros)
  - Reusability, single deployment, maintenability, and extensibility
- (Cons)
  - A host system should allow a docker-based container technology
  - A host system should allow a web app (solution : legacy pilot with ssh or others supported by Radical Pilot

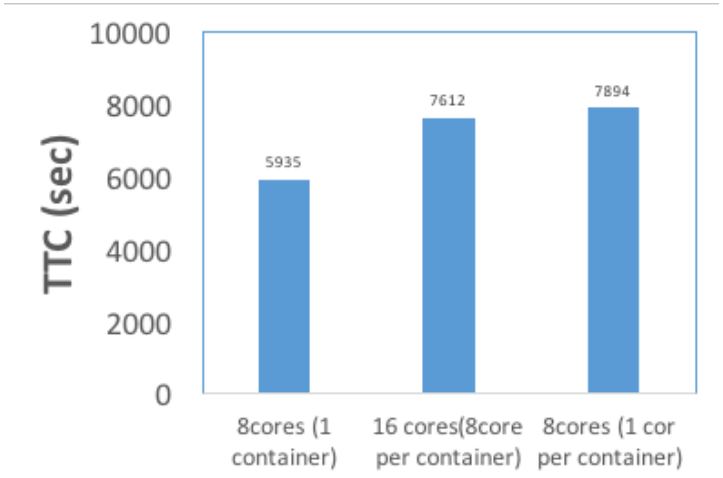
Job Execution with Docker

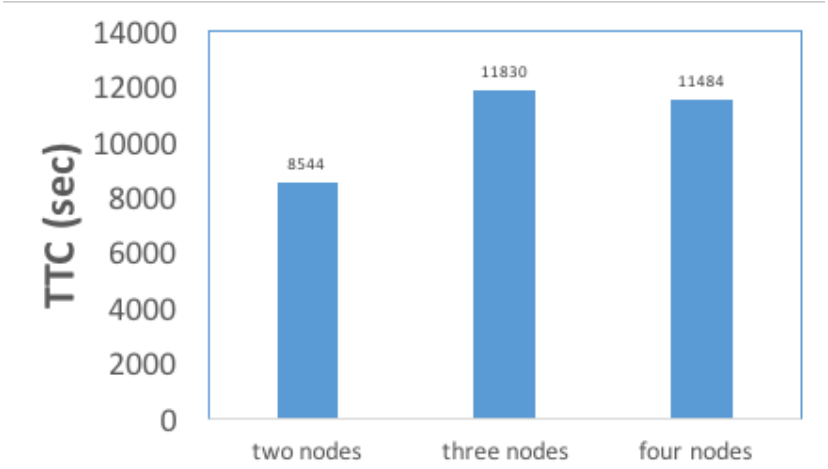
Example with the DNase-seq data analysis

- Package the target application as a docker image, along with a rest web server
- Store the docker image in the object store with SWIFT
- Start (optionally) a VM instance and initiate the service back-end with pulling an docker instance onto remote systems (Bluemix, OpenStack (Chaemeleon and JetStream)
- Start a new job, monitor and download output
- Reuse the instance by submitting another jobs
- Terminate a docker instance (optionally with VM)

Job Execution with Docker

as a scalable environment!!!





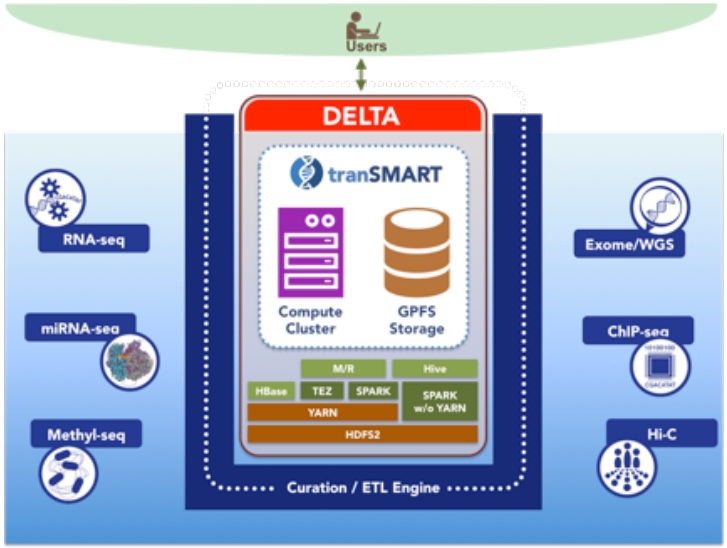
- MPI over multiple containers over multiple nodes
- Over multiple heterogeneous Clouds and HPCs
- Need to figure out the best performance setting

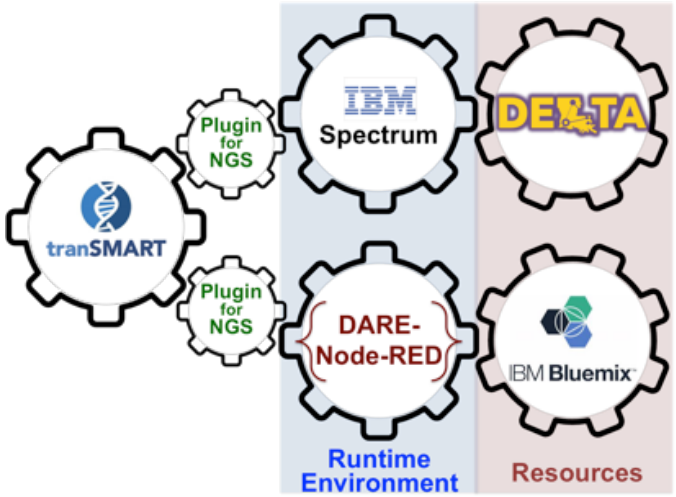
LOVE PURPLE
LIVE GOLD

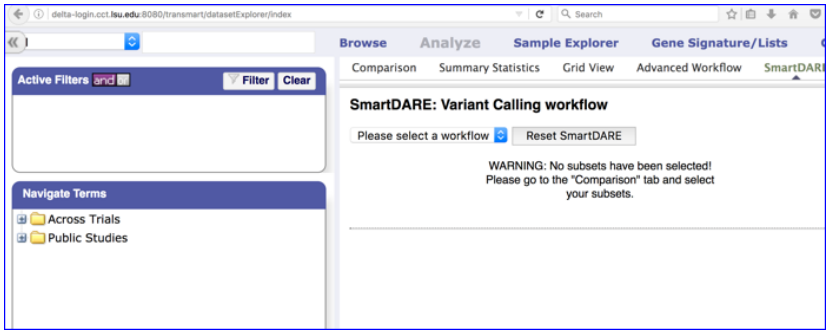
DARE Execution Scenarios

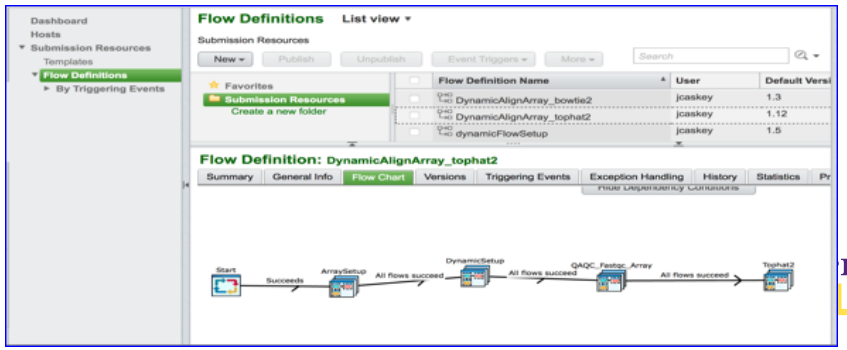
	Somatic Mutation Discovery	Transcript/Metagenome	Differential Gene Expression	DNase-seq
Analysis Objective Description	Variant calling for Cancer vs. Normal	De novo genomic sequence reconstruction	Major analysis for RNA-seq data	Deep Learning based DHS analysis for cell types
Experimental Platform	Whole Genome Seq., Exome Seq.	RNA-seq	RNA-seq	DNase-seq
On demand computing clouds	Not Yet	YES	Not Yet	In Progress
HPC	YES	YES	YES	N/A
Docker	In progress	In progress	In progress	YES

DELTA, IBM Power8 HPC, and tranSMART For Translational Research

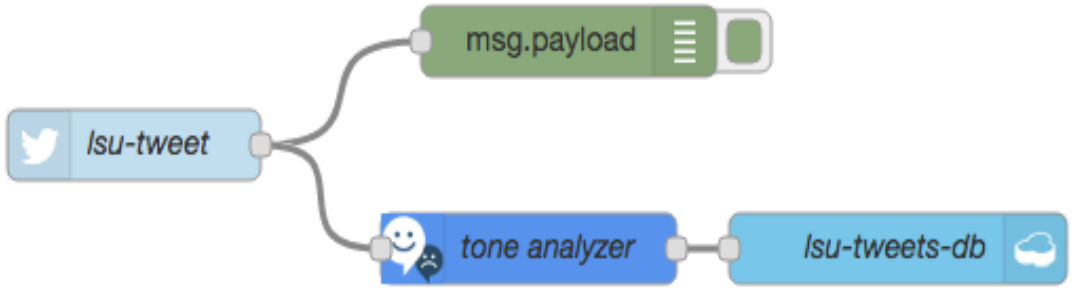


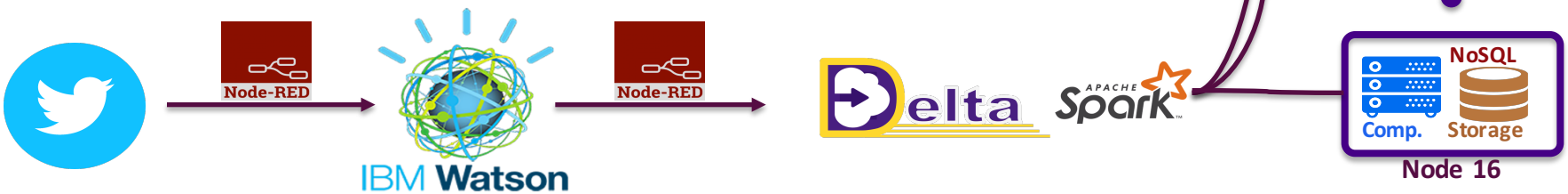






DELTA, IBM Power8 HPC, IBM Bluemix Cloud Platform and Node-RED





LOVE PURPLE
LIVE GOLD

SUMMARY

- DARE can support an optimal execution pattern over heterogeneous distributed resources by supporting a diverse set of local environments and parallel programming models
- Responsive and rich user experience web UI will be available soon
- DARE-BigNGS is good for a small size group to conduct large scale data analyses (noisy and big)

Where to go now?

- Fully automatic execution with parameter optimization (AI)
- Packaging and documentation
- Embracing modern web UI technologies
- Advanced data management system

References

1. Maddineni, Sharath et al., "Distributed application runtime environment (DARE): a standards-based middleware framework for science-gateways", *Journal of Grid Computing* 10.4 (2012): 647-664.
2. Shayan et al., "Developing A Scalable Platform For Next-Generation Sequencing Data Analytics Over Heterogeneous Clouds and HPCs : A Case for Transcriptomes and Metagenomes", Supercomputing 2016
3. Shams, S., Kim, N., Meng, X., Ha, M. T., Jha, S., Wang, Z., & Kim, J. "A Scalable Pipeline For Transcriptome Profiling Tasks With On-demand Computing Clouds". *HiCOMB2016, IEEE IPDPS* (2016)
4. Kim, J. et al (2016) "A Model for Enabling Scalable Multi-Omics Sequencing Data Analytics with TranSMART", the 3rd place prize for the poster presentation, tranSMART annual meeting

Acknowledgement

- Shantenu Jha (RADICAL, Rutgers U.)
- Zhong Wang (JGI)
- Gus Kousoulas (LBRN and COBRE)
- Shayan Shams, Mohammad M. Jalalzai, and former students
- Seung-Jong Park and Ram Ramanujam

❖NIH (COBRE, INBRE)
❖IBM (Frank Lee and IBM staffs)
❖Local funding : NuPotential INC, and many
❖Amazon EC2, IBM Bluemix, Chameleon, and ALCF

LOVE PURPLE
LIVE GOLD