Convergence of HPC & Bigdata

Yutong Lu

yutong.lu@nscc-gz.cn
National Supercomputer Center in Guangzhou
China

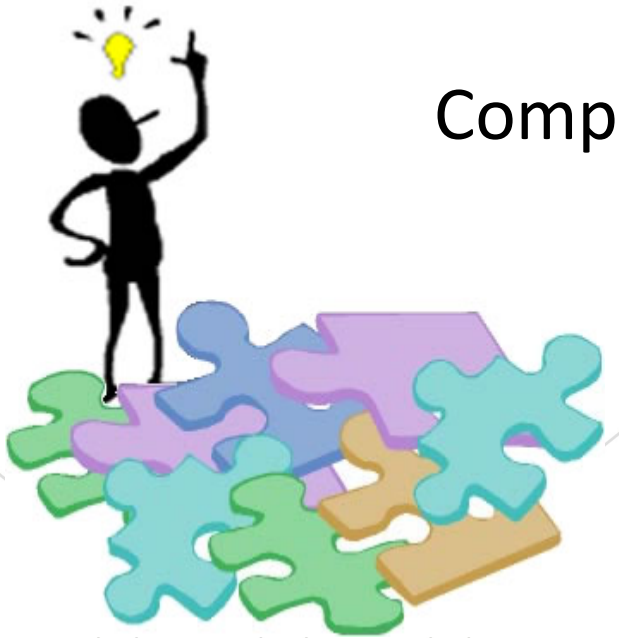
Big compute V.S. Big data

Compute-intensive

vs.

Data-intensive

Idea



Compute-bound

vs.

IO-bound

Problem

Data-Intensive Simulation and Analytics

- **Tasks involving sufficient data volumes and algorithmic complexity to require HPC resources**
  - Established (simulation) or newer (analytics) methods
  - Structured data, unstructured data, or both
  - Regular (e.g., Hadoop) or irregular (e.g., graph) patterns
  - Government, industry, or academia
  - Upward extensions of commercial business problems
  - Accumulated results of iterative problem-solving methods

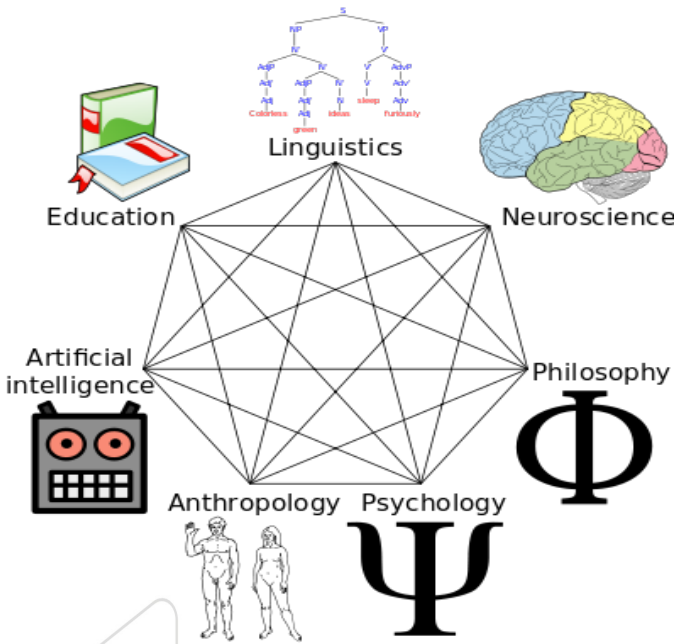
Challenges of Computing/Throughput

- **Features**
  - Massive volume, weak relationship, poor access locality
  - Various and complex data structure
  - Produced consistently and fast, analysis in situ
  - Extracting valuable information and knowledge from huge data

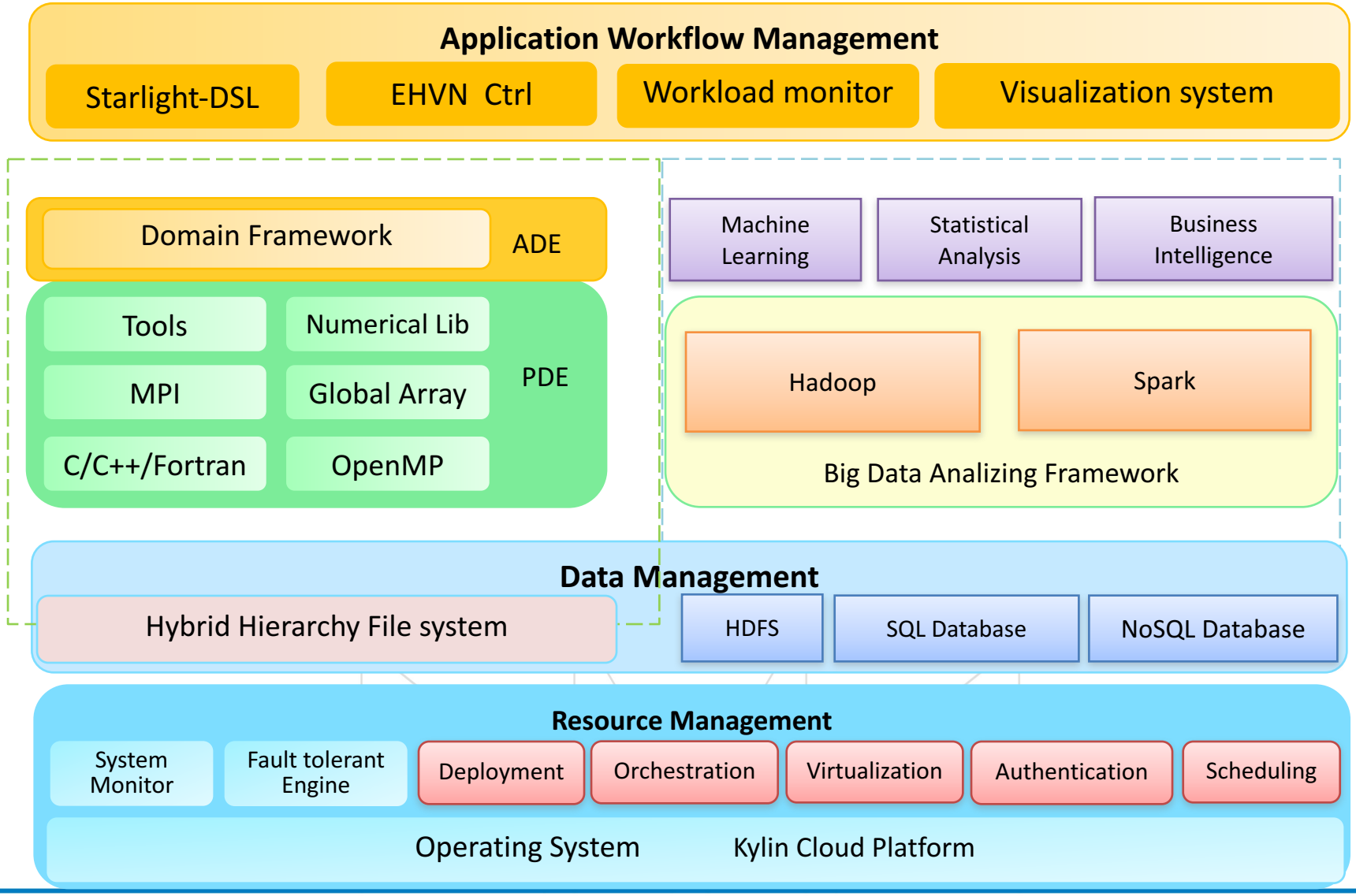
- **Computation**
  - Performance
- **Communication**
  - Scalability
- **Capacity**
  - Throughput
- **Usability**
  - Flexibility
- **Reliability**
  - Resilience

Tianhe-2’s Enhanced Designs for Big Data

- **Communication enhancement**
  - TH-Express2+
  - TCP/IP over TH-Express2+
- **I/O enhancement**
  - Hybrid Storage Structure
  - Hybrid Hierarchy File System & Data management
- **Software Stack enhancement**
  - KylinCloud Platform
  - Starlight software stack



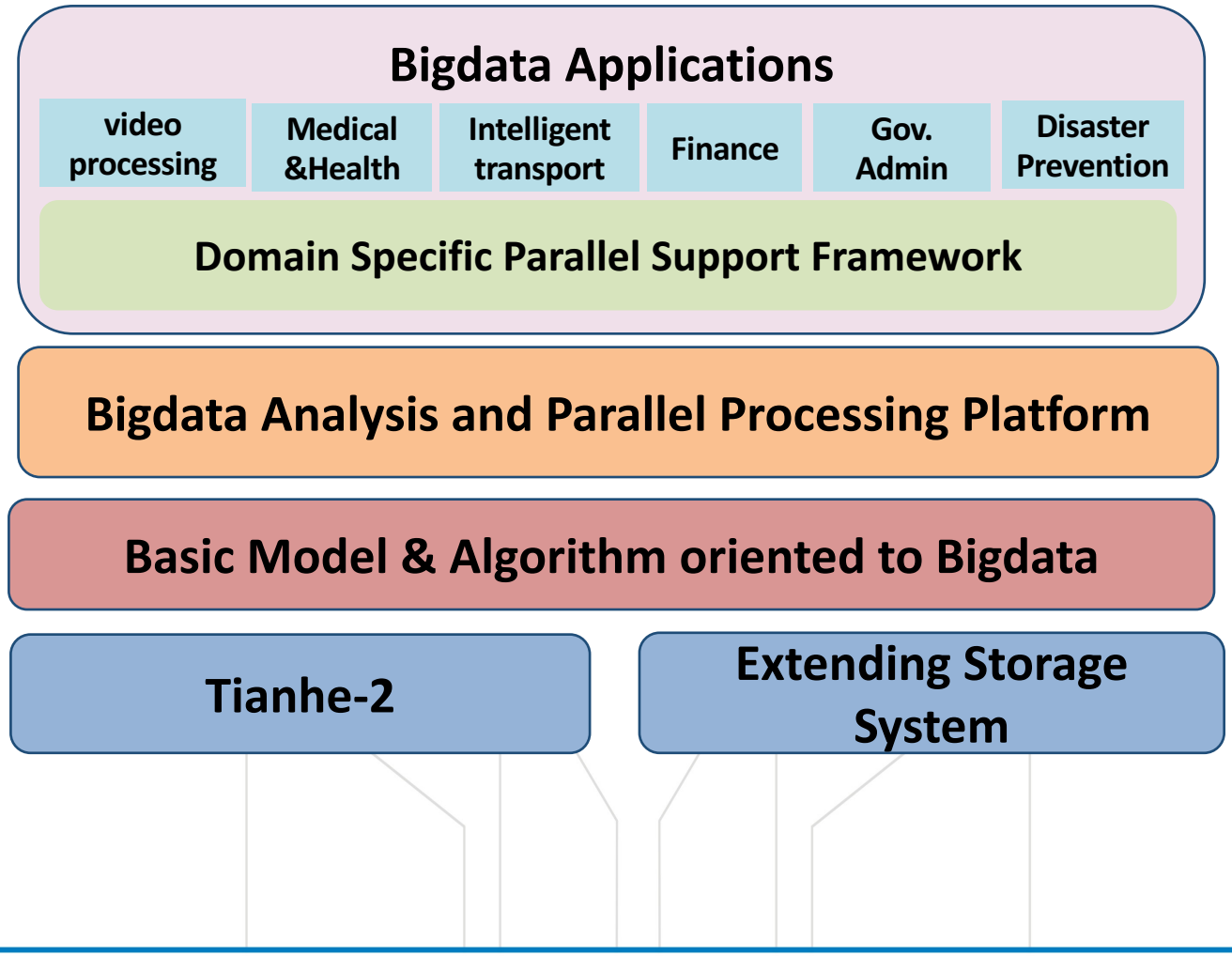
TH-Starlight Software Stack



Dealing with the Bottleneck

- How does the application consume the resource?
- Managing the data come from different sources
- Improving network performance
- Task scheduling and Workload management

Application Driven



Convergence of HPC and Bigdata

Application Driven

- Architecture and Storage system
- Application and System software
- Computation-model and Data-model
- Multi-model of Application Workflow
- Local-area and Wide-area Environment