CERIAS

The Center for Education and Research in Information Assurance and Security

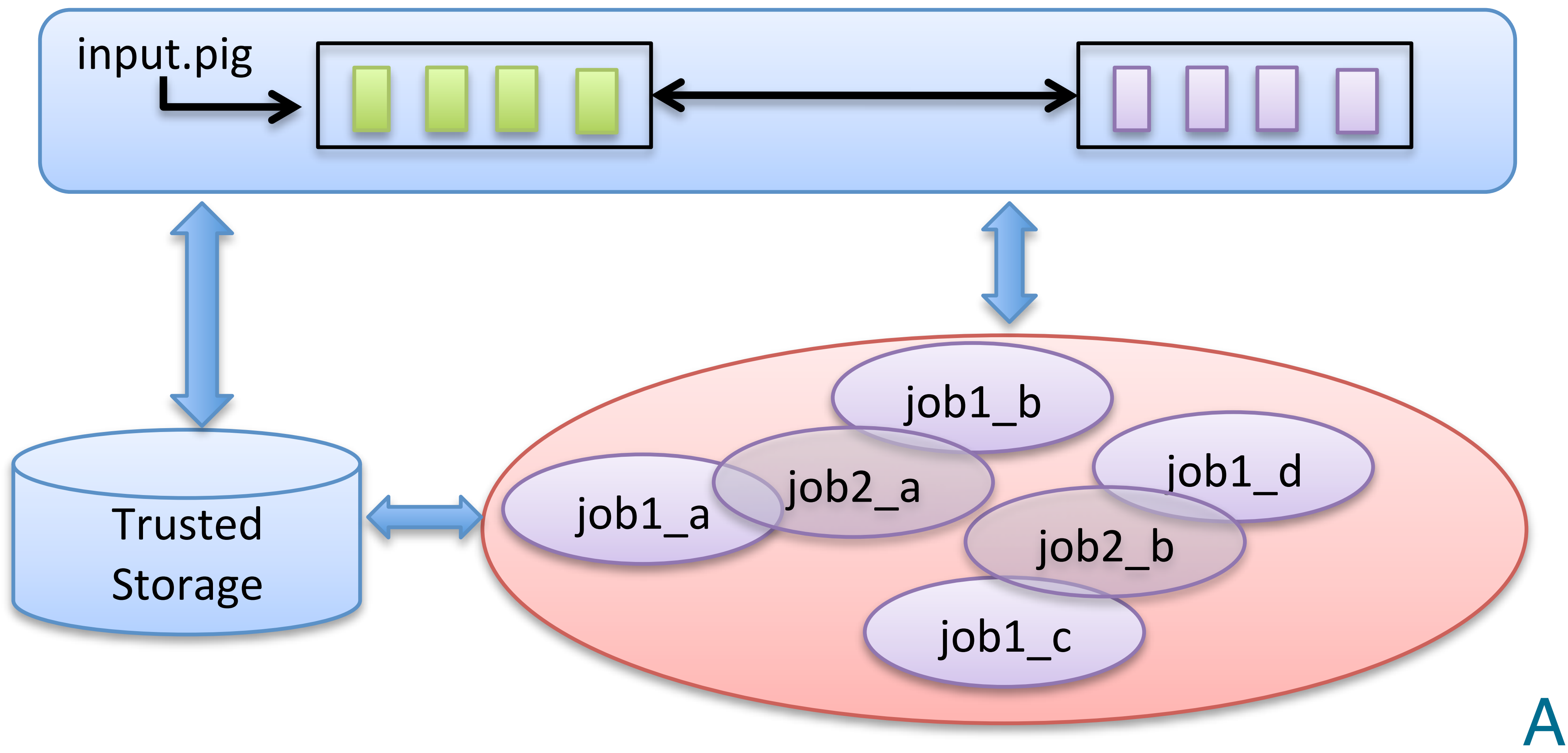
Secure Big Data Computations in the Cloud

Julian Stephen and Patrick Eugster
Department of Computer Science, Purdue University

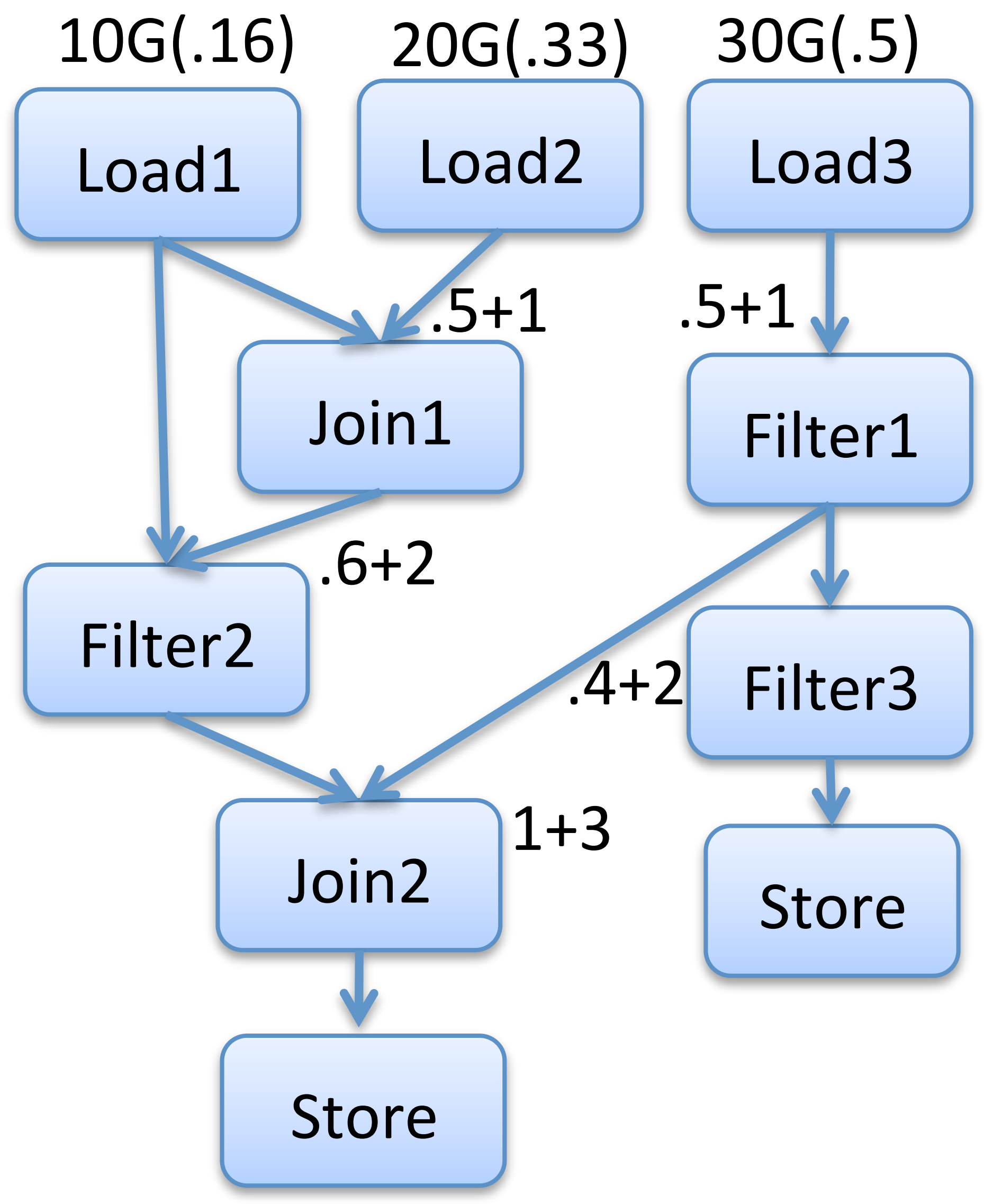
Security in the Cloud

- Communication centric:
  - Focus on messages exchanged between machines.
  - Firewalls, anti-virus, etc.
- Data centric:
  - Focus on data at rest.
  - Encryption, access control, etc.
- Computation centric:
  - Focus on computations generating correct output.
  - Byzantine fault tolerant replication, output verifiability.
- Solutions overlap, need to secure all three fronts.

Architecture



Data Flow Analysis



Identify nodes in the graph, where consensus gives maximum dividend.

Algorithm:

```
for i = 1 to #consensus points
   for each node n,
         v = ipratio(n) + dist(n);
   end for
   mark node with highest(v);
end for
```

ipratio(node)
fraction of total input processed by node at its level.

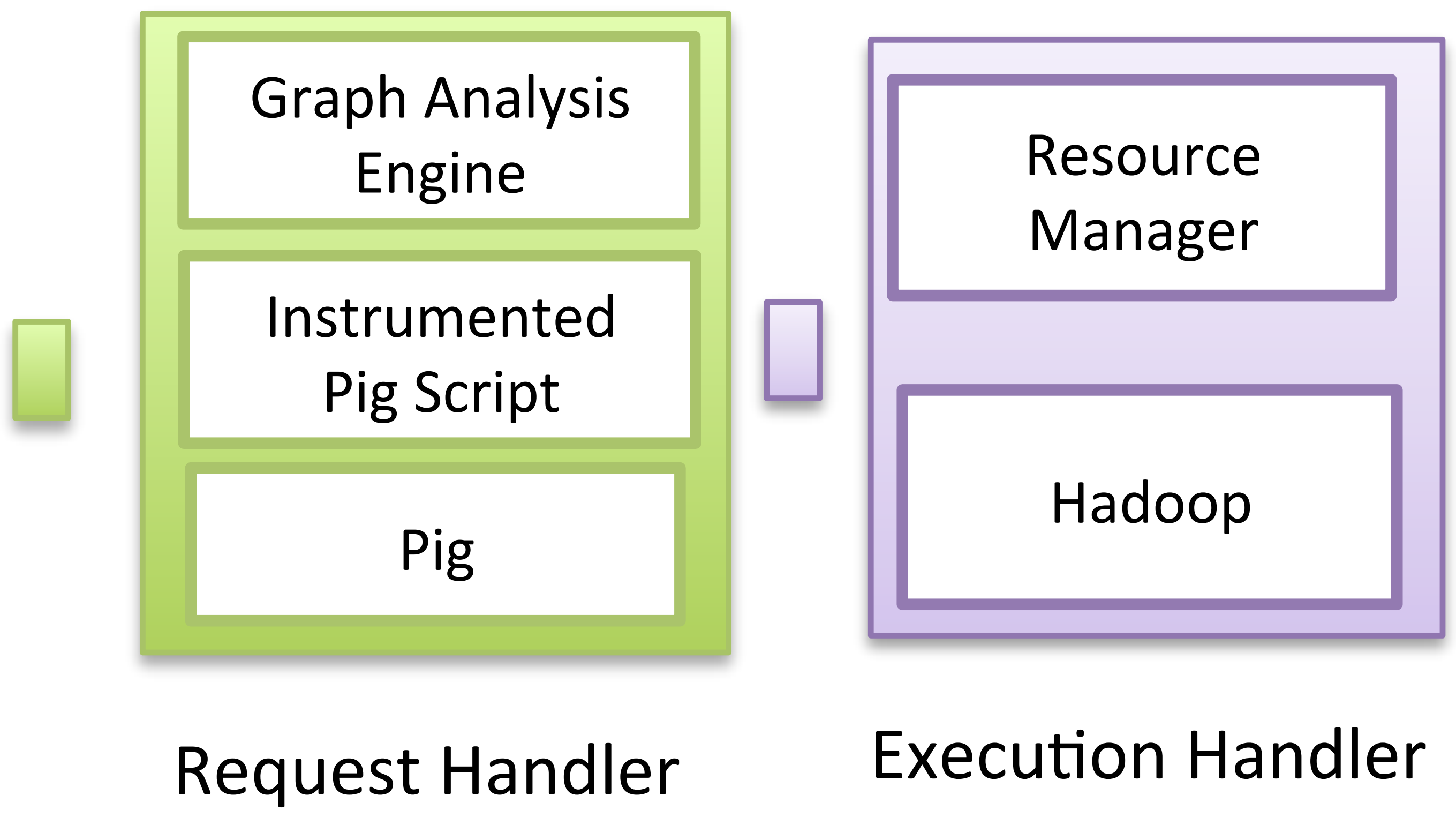
dist(node)
distance of this node from closest marked node.

Research Goals

- Secure Computations: Tolerate benign/malign faults in computing processes using byzantine replication.
- Minimize Overhead: Limit overhead caused by comparisons and re-computations.
- Attribution: Identify potentially faulty components.
- Portability: Work on multiple clouds, with different infrastructure.

BFT Replication Challenges.

- No monolithic server: Single client request executed by multiple nodes (eg Mapreduce).
- Size of data: Comparisons and re-computations highly expensive.



Agreement

- The instrumented pig script creates output digests at consensus points.
- Execution handler ensures agreement among all digests.

Attribution

- Run multiple jobs such that computation nodes overlap.
- Increase suspicion level of nodes that return faulty results.

Portability

- Can run on top of any data flow based big data analysis language - (DryadLINQ, PigLatin)

Implementation

- Modifications to pig interpreter to instrument pig scripts for creating output digests.
- Modifications to Hadoop resource allocator to enforce replica placement and node overlap.

Future Work

- Homomorphic encryptions can be used to protect against malicious computations leaking information.
- Runtime statistics provide more accurate information to identify better consensus points.