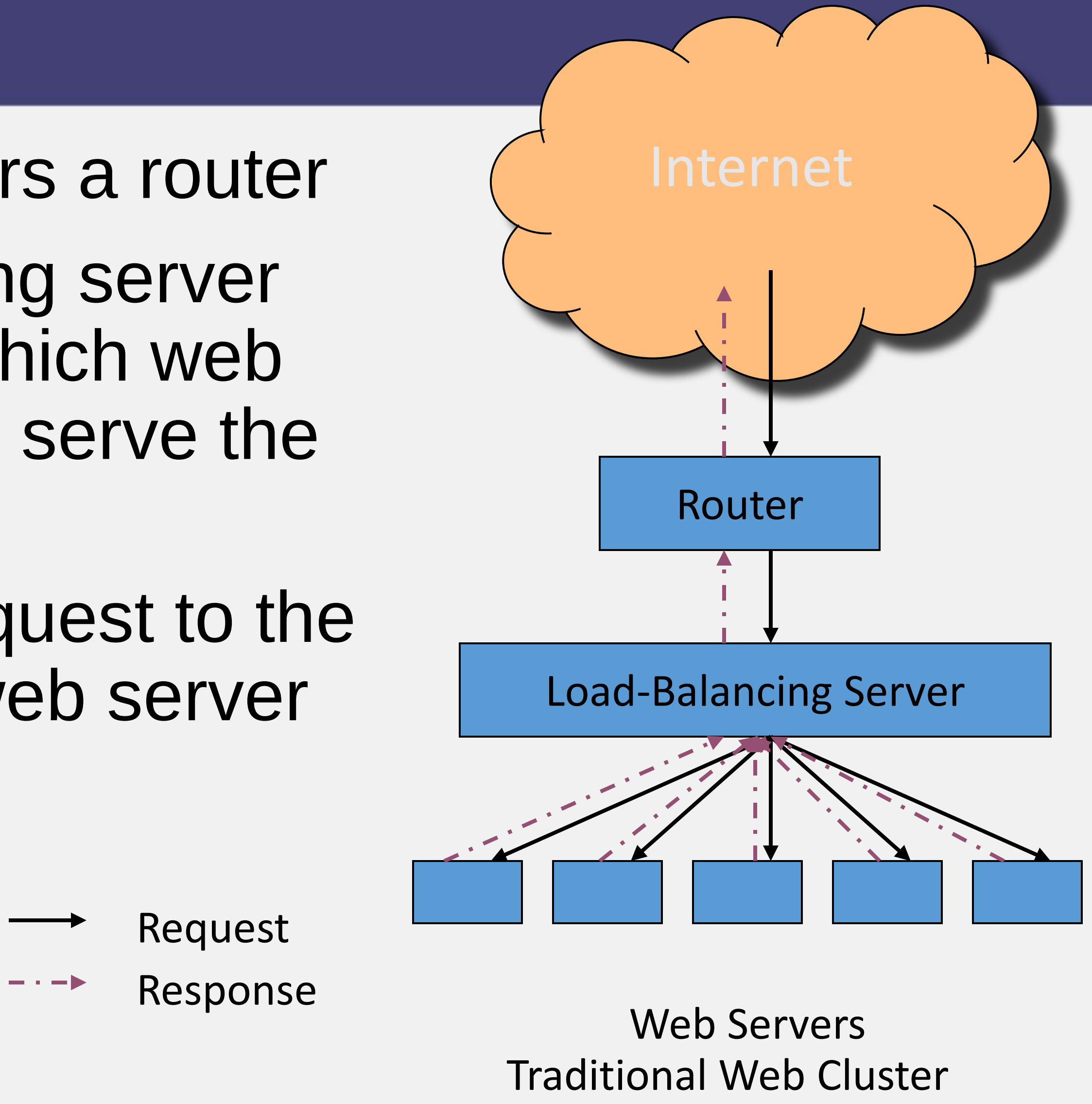
CLOUD COMPUTING APPLICATIONS

LOAD BALANCER INTRO

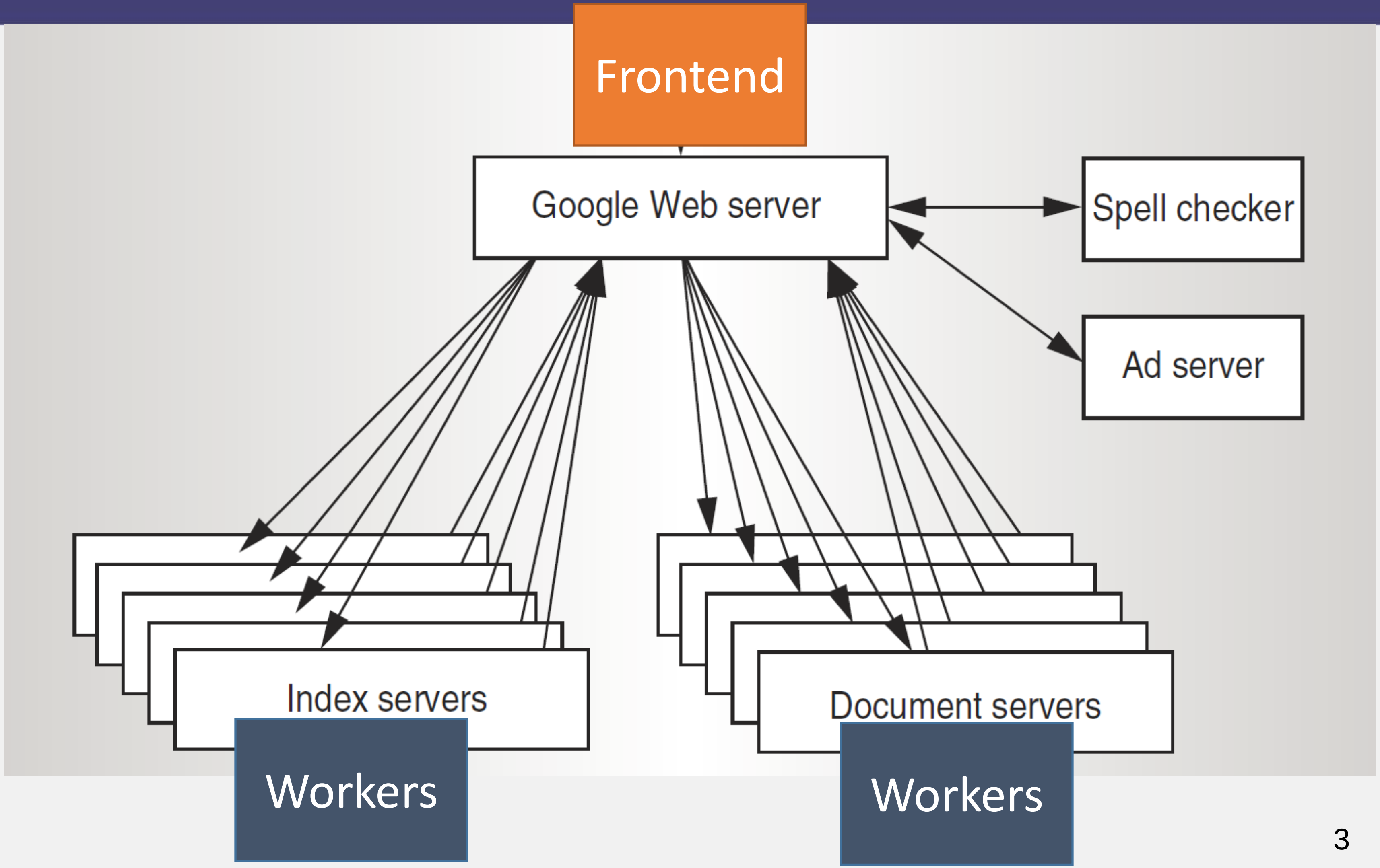
Prof. Roy Campbell

Introduction to Load Balancing

- Request enters a router
- Load balancing server determines which web server should serve the request
- Sends the request to the appropriate web server

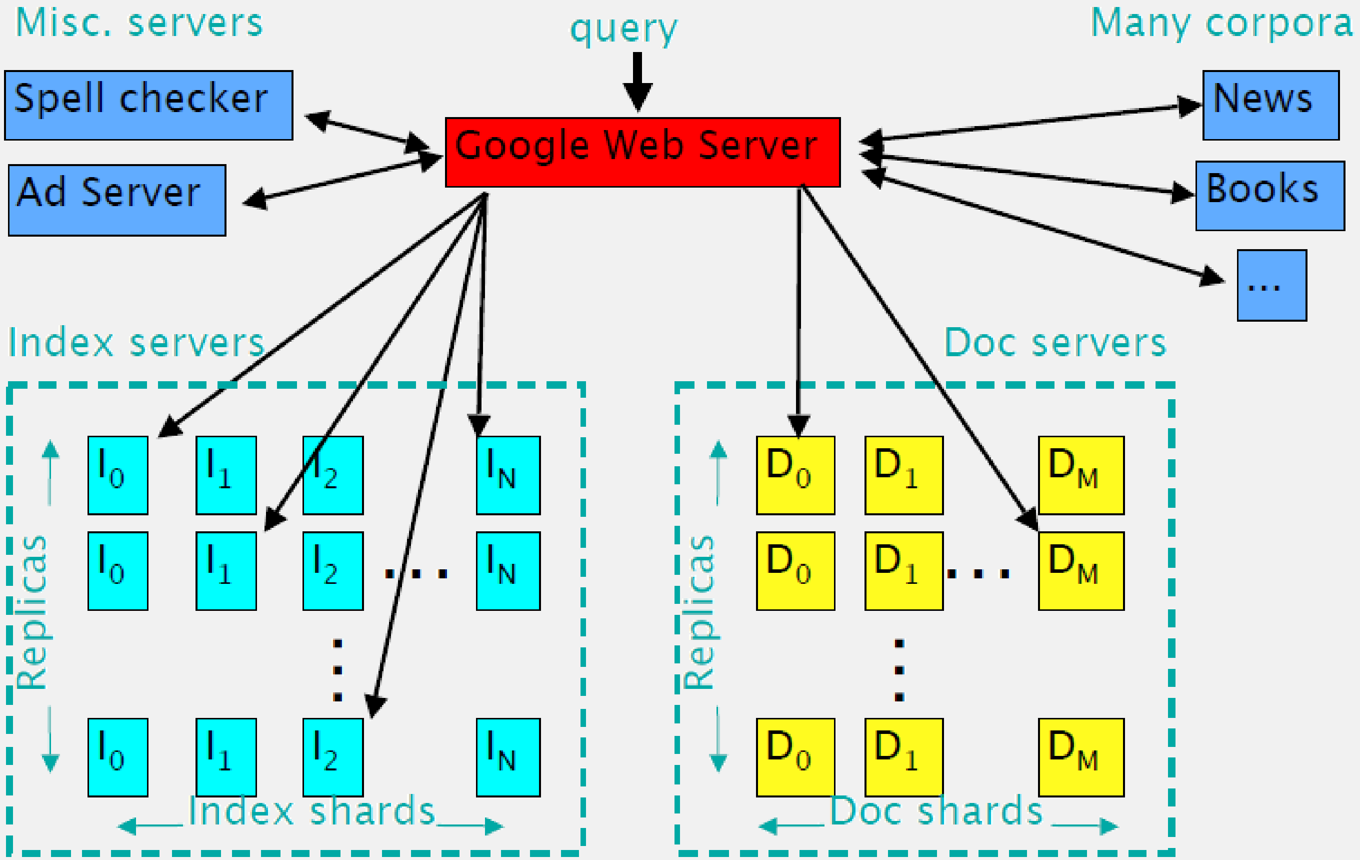


Web Search for a planet: The Google Cluster Architecture (2003)



Google: A Behind-the-Scenes Tour

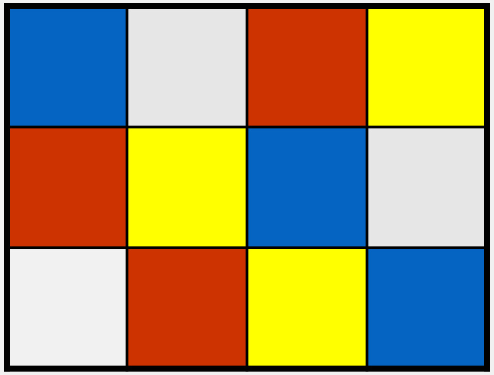
Google Query Serving Infrastructure



Elapsed time: 0.25s, machines involved: 1000s+

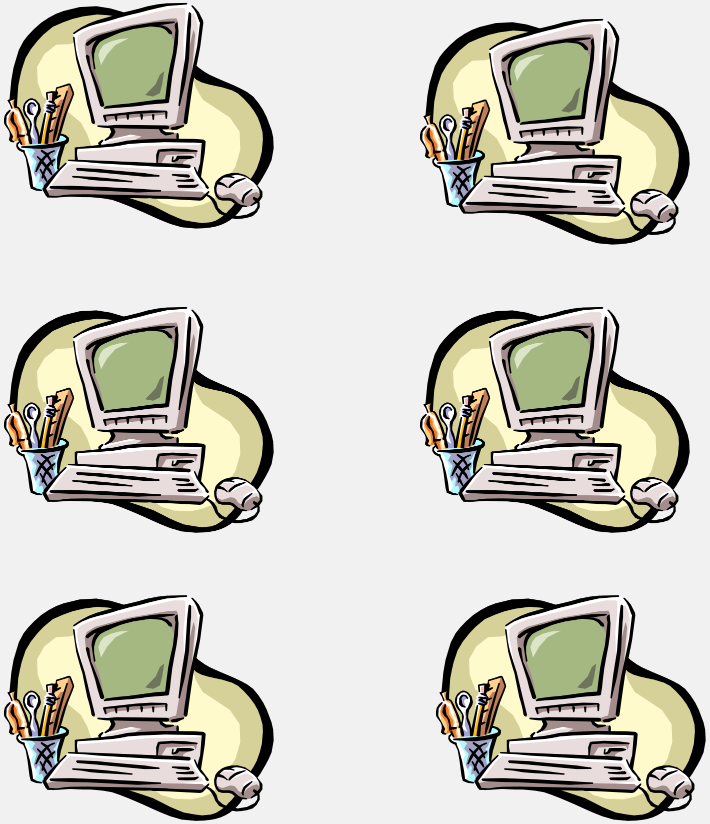
How do we split up information?

Content



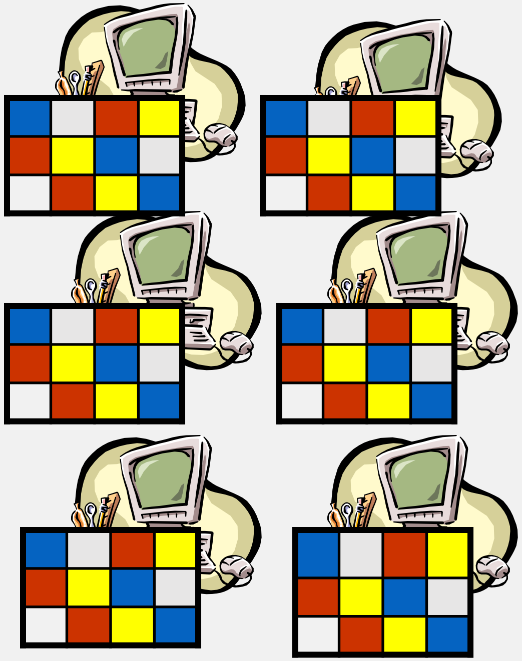
?

Server Farm

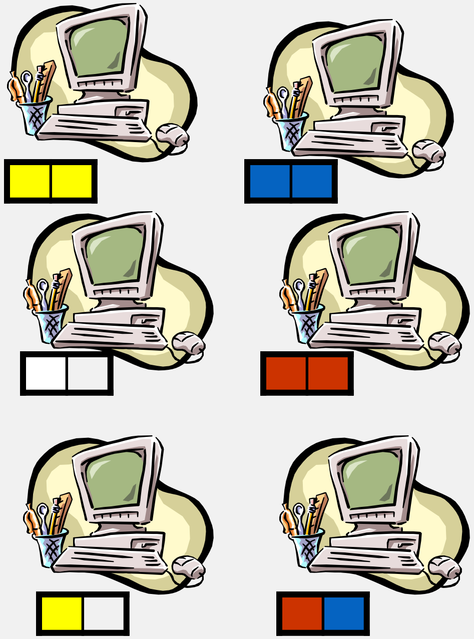


Information Strategies

Replication



Partition



Load Balancing Approaches

File Distribution	Routing
Content/Locality Aware	DNS Server
Size Aware	Centralized Router
Workload Aware	Distributed Dispatcher

Issues

- Efficiently processing requests with optimizations for load balancing
  - Send and process requests to a web server that has files in cache
  - Send and process requests to a web server with the least amount of requests
  - Send and process requests to a web server determined by the size of the request