Hadoop 簡介

- Google 中 MapReduce 概念的實現
- Apache Hadoop is a Free Java software framework ，創始人為 Doug Cutting
- Hadoop 為分散式文件系統 (DFS) ，設計來用在大型 Cluster 上執行分散式應用的框架

Hadoop 與 Google 的對應

Develop Group	Google	Apache
Sponsor	Google	**Yahoo, Amazon**
Algorithm Method	MapReduce	MapReduce
Resource	open document	**open source**
File System (MapReduce)	GFS	**HDFS**
Storage System (for structure data)	big-table	**Hbase**
Search Engine	Google	Nutch
OS	Linux	Linux / GPL

Hadoop 特點

Hadoop 包含兩大項：

- HDFS
  - architecture
  - 寫入
  - 讀取
  - 備份策略
  - 穩定性
- MapReduce

Hadoop 的特點：

- 把 MapReduce、 HDFS 及 JobTracker、 TaskTracker 合在一起
- 避免不必要的檔案搬運
- 排程器根據檔案系統現況，把合適的計算丟到特定資料節點。

Hadoop 與雲端運算

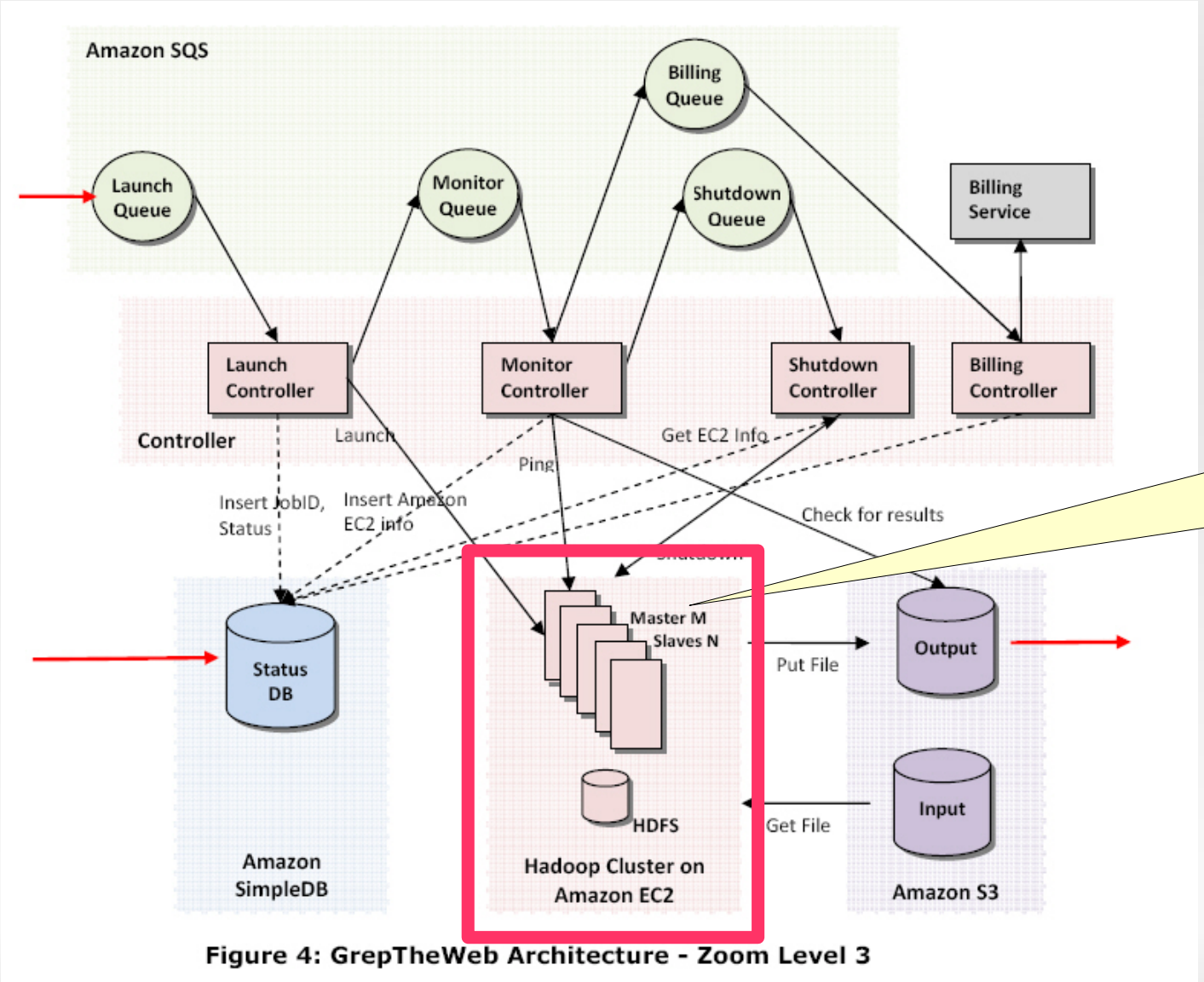

Figure 4: GrepTheWeb Architecture - Zoom Level 3

左圖是 Amazon 的雲端運算架構 (Cloud Architecture)

Amazon EC2 用 Hadoop 打造

用 DRBL 佈署 Hadoop

- 只要安裝一台 DRBL Server
- 在 DRBL Server 上設定好 Hadoop
- 每台 DRBL Client 就可充當 MapReduce 的計算節點與儲存節點。
- 缺點：
  - 如果要當儲存節點必須使用 DRBL Client 的硬碟，與 DRBL 無碟的原始精神略有落差。
  - 可改用 AoE (ATA over Ethernet) 提供儲存空間。

DRBL 佈署 Hadoop 展示

- DEMO 1:
  - 不使用 DRBL Client 硬碟，只貢獻出 /temp 當 HDFS 。缺點 : 不能儲存大量資料。
- DEMO 2:
  - 使用 DRBL Client 硬碟 (D:)
- DEMO 3:
  - MapReduce 實際範例 : wordcount