GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 18: PARALLEL COMPUTING

Francesco Mosconi
DAT5 SF // May 7th, 2014

LAST TIMES:

- NAÏVE BAYES
- PLOT.LY
- HISTORY OF MACHINE LEARNING

QUESTIONS?

AGENDA

I. INTRO TO CLUSTER COMPUTING

II. LAB: PARALLEL PARAMETER GRID SEARCH

I. CLUSTER COMPUTING WHAT, WHY, HOW

Q: What is a cluster?

Q: What is a cluster?
A: A **computer cluster** consists of a set of **connected computers** that **work together** so that in many respects they can be viewed as a **single system**.

http://en.wikipedia.org/wiki/Computer_cluster

Q: Why use a cluster?

Q: Why use a cluster?
A: General and Specific reasons

Q: Why use a cluster?

A: General:

- Lower Cost: pay-as-you-need
- Elasticity: add and remove resources
- Availability: launch jobs through API

Q: Why use a cluster?

A: General:

- Lower Cost: pay-as-you-need
- Elasticity: add and remove resources
- Availability: launch jobs through API

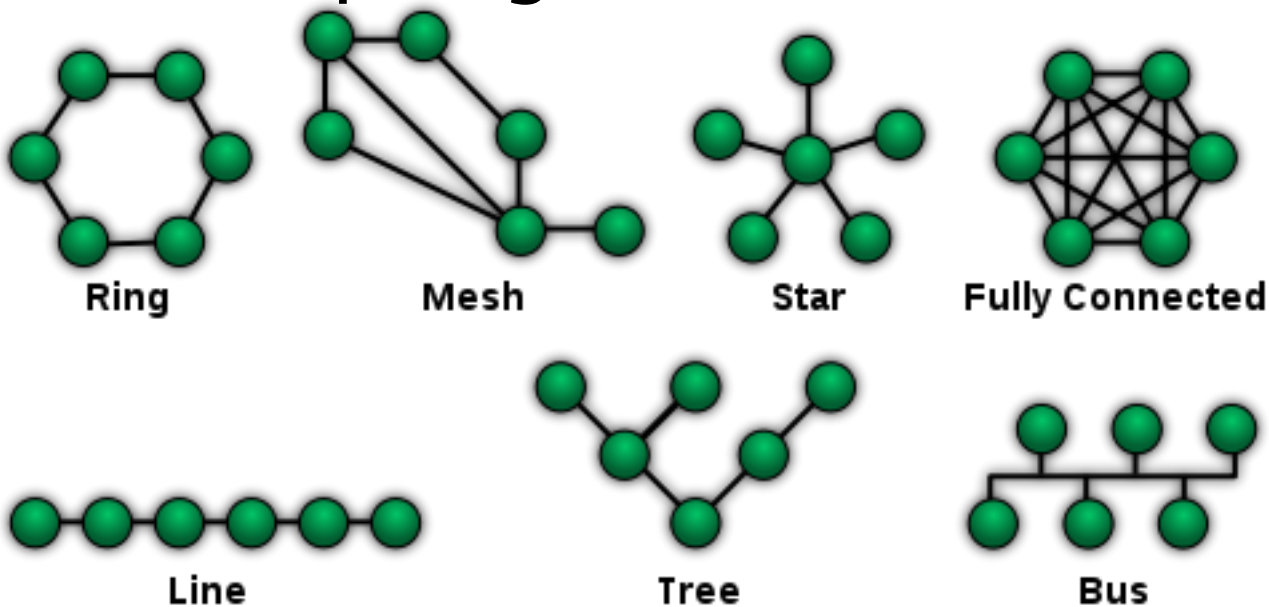
Specific to Data Science:

- Distributed Map Reduce
- Data doesn't fit in memory

Q: How is a cluster formed?

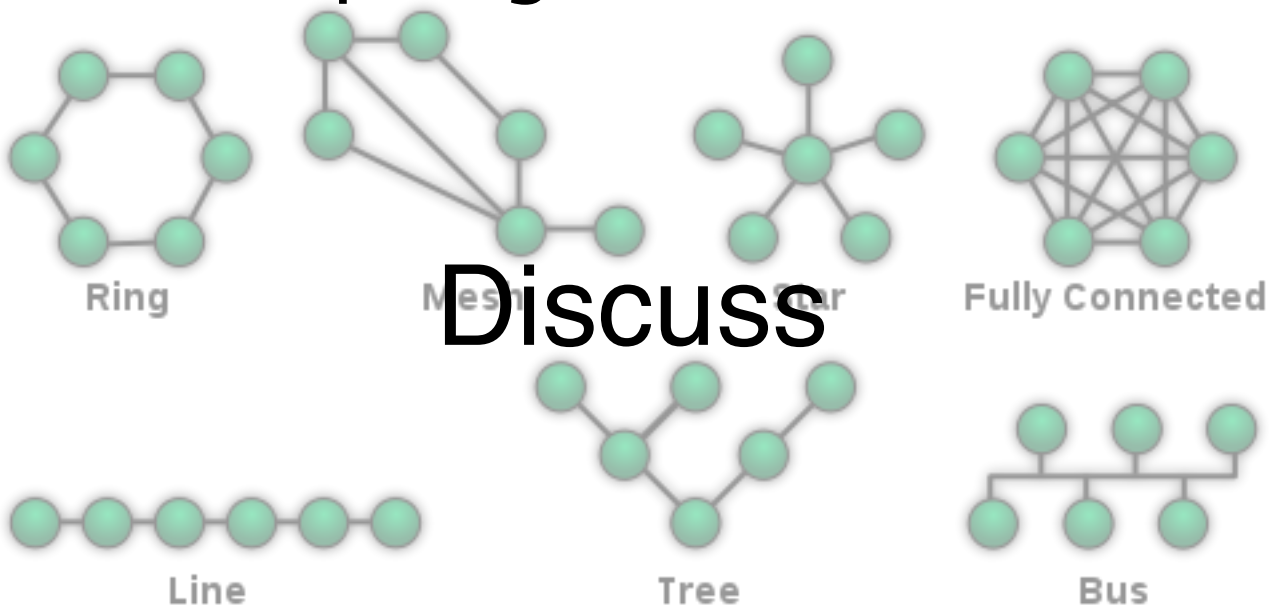
Q: How is a cluster formed?
A: Different topologies

Q: How is a cluster formed?
A: Different topologies



http://en.wikipedia.org/wiki/Computer_cluster

Q: How is a cluster formed?
A: Different topologies



http://en.wikipedia.org/wiki/Computer_cluster

Q: Benefits of Start Topology

Q: Benefits of Start Topology
A:

- Better performance: max 3 dev and 2 links
- Isolation: non-centralized failure no effect
- Centralization: control, fault detection
- Easy Install and config

Q: Features of a cluster?

Q: Features of a star cluster?
A:

- Dispatch jobs to whole cluster
- Control individual nodes
- Shared memory

II. LAB: IPYTHON PARALLEL