Exploring the Effect of Heterogeneity in Distributed Systems

Brighten Godfrey

Work with Ion Stoica

OASIS Retreat
January 11, 2005

Exploring the Effect and Cause of Heterogeneity in Distributed Systems

- **Motivation** High level of heterogeneity exists among internet routers, peer-to-peer systems, etc.
- **Question** Two systems A, B with same total "capacity", but
  - Nodes of A have equal capacities
  - Nodes of B have unequal capacities

  When does A or B perform better?
- **High level goals**
  - Understand effect of various levels of heterogeneity in distributed systems
  - ... and therefore why certain distributions arise
  - Develop general techniques to handle and exploit heterogeneity

This talk

- Will describe early stages of work; your comments are appreciated
- Outline

  1. Some examples
  2. Quantifying "heterogeneity" and its effect
  3. What happens when there is one resource in the system identical at all nodes? (Some preliminary results)
  4. What happens when nodes have diverse resources?

This talk

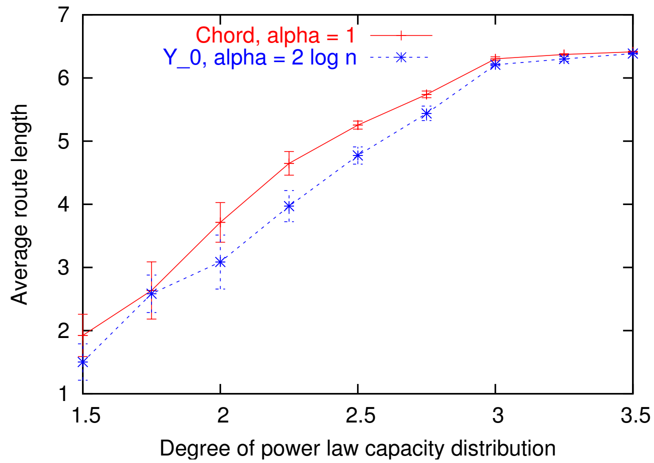
WARNING

THIS TALK HAS BEEN RATED

EF13 **EXPLICIT FORMALISM**

Examples

- **Overlay routing**
  - Capacity is available bandwith at overlay node
  - Equal capacities $\Longrightarrow O(\log n)$ hops to route;
    One node has huge capacity $\Longrightarrow O(1)$ hops to route
  - Simple techniques can take advantage of heterogeneity in DHTs...



Route length vs. capacity distribution in a 16,384-node system [Godfrey, Stoica '05].

- **Load balancing** in DHTs [SGLKS'04]: significantly better balance in real-world Gnutella capacity distribution vs. homogeneous

Examples

- **Heterogeneity might not always help...**
- e.g. ten-process simulation; processes synchronize after each time period; all take equal computation
- Obviously, ten 1000 MHz processors better than nine 1100 MHz processors and one 100 MHz processor

Defining heterogeneity

- *Capacity vector* $C = (c_1, \ldots, c_n)$:

$$c_1 \geq \cdots \geq c_n \geq 0 \qquad \text{and} \qquad \sum_i c_i = n.$$

- *Majorization* partial order: C' *majorizes* C, written $C' \succeq C$, when for any $k \in \{1, \ldots, n\}$,

$$\sum_{i=1}^{k} c'_i \geq \sum_{i=1}^{k} c_i.$$

- Intuitively...
  - C' is "more heterogeneous" than C, or
  - C' is "more centralized" than C

Defining heterogeneity: Why majorization?

- First arose in economics to compare income distributions
- Bottom $\perp = (1, \ldots, 1)$ is homogeneous distributed system
- Top $\top = (n, 0, \ldots, 0)$ is centralized system
- Going from C to $C' \succeq C$, “the rich get richer”:
  $C' \succeq C$ iff one can produce C' by starting with C and performing a sequence of *capacity transfers* from lower- to higher-capacity nodes.
- $C' \succeq C$ implies $\mathrm{var}(C') \geq \mathrm{var}(C)$

Defining the effect of heterogeneity

Two statements to make:

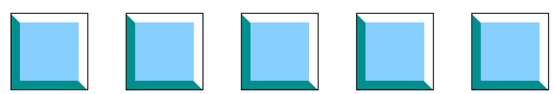
- **Average case:** "usually, heterogeneity improves performance" — future work...
- **Worst case:** "heterogeneity sometimes hurts, but never much"
  - $OPT(C, O)$ is the cost of the optimal solution for capacities C and arbitrary problem-dependent workload O
  - e.g. $OPT(C, O) =$ time for processors C to complete jobs O under best possible job schedule
  - *Price of diversity* (PoD) of a problem is

$$\sup_{O, C, C' \colon C' \succeq C} \frac{OPT(C', O)}{OPT(C, O)}.$$

  - PoD of $5/4$ says that for any systems C and $C' \succeq C$, C' can handle any workload with cost at most 25% higher than C.

Why heterogeneity might help

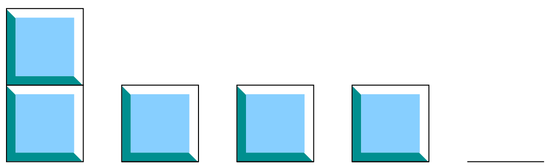
- Recall: can produce C' from C through capacity transfers to higher-capacity nodes
- Put restriction on transfers: each step moves the *full* capacity of one node to another.



- Then at each step, gainer can simulate whatever work loser was doing, and also might take advantage of some *economy of scale*

Why heterogeneity might help

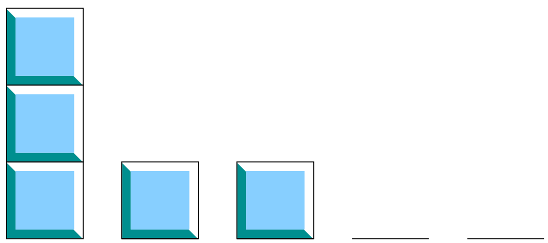
- Recall: can produce C' from C through capacity transfers to higher-capacity nodes
- Put restriction on transfers: each step moves the *full* capacity of one node to another.



- Then at each step, gainer can simulate whatever work loser was doing, and also might take advantage of some *economy of scale*

Why heterogeneity might help

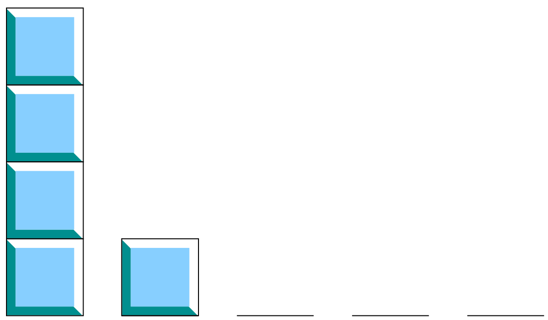
- Recall: can produce C' from C through capacity transfers to higher-capacity nodes
- Put restriction on transfers: each step moves the *full* capacity of one node to another.



- Then at each step, gainer can simulate whatever work loser was doing, and also might take advantage of some *economy of scale*

Why heterogeneity might help

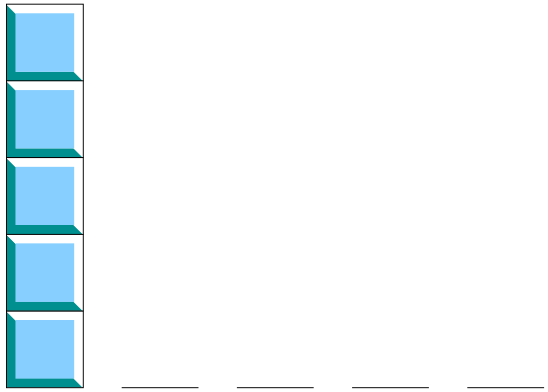
- Recall: can produce C' from C through capacity transfers to higher-capacity nodes
- Put restriction on transfers: each step moves the *full* capacity of one node to another.



- Then at each step, gainer can simulate whatever work loser was doing, and also might take advantage of some *economy of scale*

Why heterogeneity might help

- Recall: can produce C' from C through capacity transfers to higher-capacity nodes
- Put restriction on transfers: each step moves the *full* capacity of one node to another.



- Then at each step, gainer can simulate whatever work loser was doing, and also might take advantage of some *economy of scale*

Questions intuition doesn't answer

1. What if we remove the "whole-capacity transfer" restriction?
2. What if one unit of capacity on machine x is not equivalent to one unit on machine y?

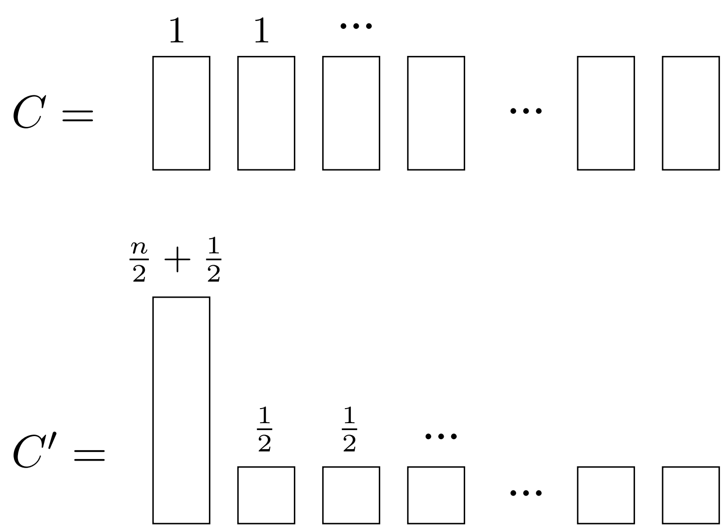
Preliminary results: Simulation lemma

- **Simulation lemma**: For any $C' \succeq C$, the C'-nodes can simulate the C-nodes with no node overextending itself by a factor > 2.

  More formally, $\exists$ an assignment $f : \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$ of the C-nodes to the C'-nodes such that $\forall i$,

$$\sum_{j \in f^{-1}(i)} c_j \leq 2c_i'.$$

- Can't do any better than a factor of 2:



Preliminary results: Simulation lemma

- Yields upper bound of 2 on Price of Diversity of minimum makespan scheduling on related machines
  - Capacity is processor speed; assign set of jobs of various lengths to processors, minimizing time last processor finishes
  - Matching lower bound of 2 on PoD (same as previous example)
- Also yields upper bound of 2 for scheduling with various other load balance metrics (probably not tight)
  - Average job completion time [Karp]
  - L_p norm of completion times, rather than makespan

Preliminary results: Building Graphs

- Also yields bound for a Graph Construction problem:
  - Given degree bounds $c_1, \ldots, c_n$, construct a graph whose nodes have degrees $\leq kc_1, \ldots, kc_n$
  - Bicriteria optimization: minimize k (degree) and diameter of graph
  - PoD bound of essentially $(2, 1)$ from Simulation Lemma
- Simulation Lemma seems not well suited to cases when capacities define hard constraints
- A different technique shows bound of $(1, 2)$ on Graph Construction
  - Restricted to trees, PoC is $(1, 1)$
  - The best tree's diameter is at most twice that of the best graph for given degree bounds

Summary so far

- If capacity on machine x or machine y is essentially equivalent...
  - Expected result: heterogeneity is generally an advantage — performance will never get much worse, and usually will improve due to economy of scale
  - Still lots of work to be done: tighten existing bounds; price of diversity of wider range of problems/distributed systems; average-case analysis
- But capacity may have different "attributes"...
  - Locality
  - Time of availability (e.g. uncorrelated failures)
  - Security vulnerabilities
  - (other suggestions?)

  What happens then?

Tradeoffs

- If system can benefit from both
  - availability of such different attributes, and
  - "economy of scale" due to increased heterogeneity/centralization,

  then we have a tradeoff between distribution and centralization.

- Example 1: Facility Location
  - Given set of customers and potential facility sites, decide where to build facilities
  - Cost depends on (1) number of facilities built and (2) the distance from each customer to its nearest facility
  - Removing (1) or (2) would result in most distributed or most centralized systems, respectively

Tradeoffs

- Example 2: Fabrikant et al
  - Model of internet graph construction
  - Network constructed one node at a time
  - Arriving node attaches to current "graph" through bicriteria optimization of locality and graph diameter
  - Result is graph with power law degree distribution (for a wide range of parameters)

Summary

- Show that if all capacity on machines is essentially equivalent, increasing heterogeneity generally helps performance
  - Some preliminary results: Simulation Lemma etc.
- Characterize the structures that arise due to a tradeoff between the economy of scale of a centralized system and the diverse resources of a distributed system.
  - Power law distributions?
- Develop set of simple techniques to adapt distributed systems to heterogeneous situations
  - e.g. discarding low-capacity nodes