

Homework #1 (distributed April 5, 2004; due April 19, 2004)

This assignment covers the ideas behind the notion of a grid and the first model problem of resource description. These topics are covered by the lectures and readings from March 31, 2004 through April 16, 2004. Do four of the problems numbered #1 thru #6. #7 is an extra credit problem.

The completed assignment can be emailed to Professor Chien, or paper copies dropped off outside my office.

1. Utility Computing: One of two major drivers behind the notion of the grid. In this question, we will explore some of the issues in presenting computing as a utility.
 - a. Suppose you are the Chief Information Officer (CIO) for BigFortune, Inc. a major enterprise doing over \$1 Billion of revenue per year. To support this business activity, BigFortune spends \$150 million per year in internal IT expenses, approximately half for desktop computing for employees and the other half for 750 servers that directly support the business, including functions such as sales, inventory, manufacturing, shipping, financial tracking, etc. If BigGridUtility, Inc. bids to support your 750 servers at a fixed cost of \$1 million and \$60,000/server. Suppose that it also costs you \$5,000/server to purchase network connectivity from your internal intranet to BigGridUtility's premises. Plot the savings versus number of servers outsourced. What number of servers outsourced will yield the greatest savings?
 - b. In order to reduce the risk in migrating your server IT infrastructure to BigGridUtility, you decide to do this incrementally, moving 100 servers the first year, 250 the second year, and 400 servers the third year. How does this choice affect your total costs for the servers over that period?
 - c. In order for your outsourcing to BigGridUtility to be stable, they have to make a profit. Ignoring for the moment the cost of sales and relationship per customer (this is significant), what is the relative efficiency (total internal cost for BigGridUtility per server) needed to support the pricing structure indicated above and a 10% profit on each sale? (plot this as a function of # of servers)
 - d. While the above efficiencies could perhaps be achieved with better technology, more efficient organization, etc. another possible way to achieve it is with reduced licensing costs enabled by sharing resources (statistical multiplexing). However such sharing depends on the load for different servers (from this customer or from distinct ones) being uncorrelated. Suppose it turns out most of the time this works, but once every 100 utility server hours, this sharing causes an application to miss a deadline. Through careful analysis, you've figured out that each such event costs your organization \$100. How does this change your answers to parts a and b? (and the resulting financial outcomes)

2. Federation and Sharing: The second major driver behind adoption of grids is the benefits of federation and sharing of information and resources.
 - a. Consider a distributed design team for automobiles which includes expert designers from each of the 35 parts vendors as well as the lead designers for SuperBig Automobile, Inc. In this case, the parts vendors are all separate companies who in some cases compete with each other, and in other cases also sell parts to SuperBig's competitor, SuperEco Automobile, Inc. Explain the benefits of grid-based sharing of information, data, and computational resources amongst SuperBig and its suppliers. Explain in one paragraph, a number of ways in which this might benefit SuperBig's business.
 - b. Federation and sharing also presents a set of major pitfalls. First from the perspective of SuperBig and then from the perspective of the parts vendors, explain a number of ways in which data legitimately useful to support the goals in part a could be exploited to the disadvantage of the supplier of the data. What are the broad challenges in this sharing?

- c. After SuperBig solves all of its grid sharing challenges with its 35 vendors, a new computer SuperTrucks is formed by a consortium in Europe. Now SuperTrucks wants to join the grid formed by SuperBig. What issues does this raise?
 - d. Now in order to complete its designs, SuperBig uses computing resources amassed from the network of over 35 companies, using on average 100 CPU's from each company. This provides a large quantity of compute resources which is critical in the latter stages of design and planning for manufacturing for each new car model. What potential problems does the introduction of SuperTrucks into the grid present? In more general terms discuss the problems new additions to a grid represent?
3. Applications: Grids support both traditional and innovative new distributed applications. There are at least five major types of applications being pursued in large scale on grids: 3-tier web-database, distributed supercomputing, online distributed computing, and data grids.
 - a. Describe one example of each of these types of applications
 - b. For each of the example applications, give the key quantitative performance goals (these may be gigaflops, availability, gigabytes per second, or some other type.
 - c. For each of the example applications, what new grid middleware capabilities are needed to support it meeting the key performance goals. That is compared to a simple best-effort client-server infrastructure used for web browsing. In your opinion, do these capabilities exist today? (justify your answer)
 4. Grid Resource Description: Consider two resource description languages from the three: RSL, Jini, and Redline. If you choose Jini, then assume that Javaspace are the mechanism, and there are attributes of a similar style to that of RSL, but the operations are those of Javaspace.
 - a. For each of the two, give an example of a resource description for a set of machines which run Windows XP, have 1GB of main memory, and a CPU over 1Ghz.
 - b. For at least one of the two, give an example of a resource specification in one which cannot be given in the other. (can't be captured precisely) Explain why this is, and discuss the implications of the inability to express it.
 - c. Consider the following specific resource specification in RSL. What are the implications of using such a strict resource specification to find a resource in a small grid? A large grid? (What if 1000 users were to simultaneously make this request in a 1,000,000 resource grid?)

```

+(&(count=80)(memory>=1GB)(executable=sf_express)
  (resourcemanager=ico16.mcs.anl.gov:8711))
(&(count=256)(network=1GB)(executable=sf_express)
  (resourcemanager=neptune.cacr.caltech.edu:755))
(&(count=300)(memory>=1GB)(executable=sf_express)
  (resourcemanager=modi4.ncsa.edu:4000))

```

- d. Consider the following general resource specification in the RSL language. What are the implications of using such a strict resource specification to find a resource in a small grid? A large grid? (What if 1000 users were to simultaneously make this request in a 1,000,000 resource grid?)

```

+(&(count=80)(memory>=1GB)
  (resourcemanager=ANY))
(&(count=256)(network=1GB)
  (resourcemanager=ANY))
(&(count=300)(memory>=1GB)
  (resourcemanager=ANY))

```

5. Application Behavior and Resource Needs: Consider a client-server streaming video application which streams at high resolution and serves thousands of simultaneous users. These applications are

well characterized, and typically use highly optimized encoders, decoders and highly compressed transmission formats.

- a. Suppose that the requirements of a single video stream on the server side due to the encoding techniques from 10Mbps to 100Mbps (as measured in 1 second intervals) and 10 MegaOps to 100 MegaOps per second (as measured in 1 second intervals). If we host the server application on a 2,000 MegaOps server, how many customer video streams can we support? Explain your answer.
 - b. Suppose we host the streaming video server in a local co-location facility, and want to purchase sufficient bandwidth to ensure we can stream to the number of customers supportable by the server processor. How much bandwidth should we purchase? (obviously we don't want to waste money)
 - c. Consider distributed collaborative visualization which combines even greater burstiness with a requirement for low latencies in the computing and communication (the distributed collaboration is tightly coupled). How does this affect the situation?
 - d. Given the resource description languages such as RSL, how would you describe the resources needed for our streaming media application server? Our distributed collaborative visualization nodes? Do these descriptions capture the resource requirements precisely?
6. Dynamic Resource and Application Behavior: Consider that applications and resources have dynamic characteristics, and such (e.g. dynamic load or performance) are often a critical basis for choosing resources.
- a. Consider two key dynamic resource properties – Unix load factor and network bandwidth. Give examples of application scenarios where an application might want to choose amongst resources based on these dynamic properties.
 - b. Because these properties are dynamic, measurement is a challenge. Consider a grid of 100,000 devices, and assume it takes 1ms to collect a data point for load factor from each resource. If it tries as hard as it can, what is the “newest” data set a single node could collect?
 - c. If all 100,000 devices attempted to collect the “newest” data possible, what would be the result? (if the system didn't fail, what is the newest data they could collect?) Please be sure to think about contention.
 - d. Dynamic properties naturally change, so even these new measurements are subject to change. Based on the NWS paper, what types of predictions for these dynamic properties can be made accurately? What time scales? What accuracy?
 - e. Suppose for the moment that one could get accurate dynamic resource information and excellent predictions for dynamic resource properties that run for hours into the future. In order to make good use of this data and choose the best resources, what do you have to understand about the application and the resource environment to select the best resources? To decide if a move would improve the performance of a running application?
7. (extra credit) Utility Computing in Historical perspective: We have experienced four decades of increasingly decentralization in computing (first minicomputers, then PC desktops, then laptops, now PDA's and smartphones). Discuss the major drivers behind this drive to decentralization (cost of computing, desire for independent control of information and computing, mobility), have they disappeared? What are the countervailing forces that encourage Utility computing? Explain why these trends are or aren't in opposition.