

Comprehend the Planet

Timothy Roscoe, Intel Research at Berkeley, August 2005

This is a personal statement and does not necessarily reflect the views of Intel Corporation

August 15, 2005

I'm sure there will be no shortage of interesting visions of the future at this workshop. I'll start negatively, by examining the problems of formulating a grand challenge in distributed systems versus, say, genomics, before proposing an idea.

The first is *visibility*: distributed systems per se are invisible to the public – we all know how hard it is to give a systems demo. Distributed systems matter hugely, but tend to be viewed from an application standpoint. We are not natural applications designers, instead we are toolbuilders. The world-changing applications that have come out of the distributed systems community have happened because the existence of the tools has opened up unforeseen possibilities.

A second is performance: distributed systems usually do something that can always be done in a centralised way, but when distributed cleverly results in higher performance, better scalability, lower cost, whatever. Hence any grand challenge looks a little less grand from the user's standpoint: "an engineer is someone who can do for a dime what anyone can do for a dollar".

Third, relevance: when trying to think of big, really big, societally beneficial challenges, one ends up asking: what does most of the world need most right now? My answer is clean water, food, health, education, and freedom from violence in all its forms. What conceivable place could distributed systems work have in this context?

This last problem is interesting, though, because of the perspective it brings. In thinking of challenges at this scale, I personally end up really not caring about service providers, or equipment manufacturers, or privileged mobile company executives, or even the burgeoning market segment of teenage girls. I worry about the planet.

My suggested grand challenge is understanding the planet, and our effect on it, in real time, from a large number of vantage points, in a large number of dimensions.

The planet is a large system. Its climate, politics, tectonics, populations, ecology, and economy are things whose dynamics we do not understand. At the same time, there's a wealth of academic scientific work in instrumenting and understanding the planet, but such work is confined to a set of academic communities – often vertical silos. The use of computers in such communities is widespread, but has typically not been closely informed by the CS community.

A success proof-point might be to demonstrate the continuous, real-time fusion of large amounts of distributed data of different forms. For example, combining local temperature, air quality, traffic information, personal health monitoring and satellite imagery at global scale – there is increasing evidence that these are closely linked.

Like all grand challenges, there are many potential spin-off benefits: for healthcare, transport systems, etc.

Large computations are required, but in addition to the (rather limited) vision of academic Grid computing (basically batch jobs), this is about continuous processing of data by long-running services, something this community has always regarded as more interesting, more useful to the world in general, and in any case a superset of the batch-processing problem. This vision is also complementary to, separate from, but not dependent on, the sensor networks vision. Sensors will provide data. Our job is to compute on this data at global scale, and deliver meaningful results to useful places.

The call mentions clean-slate redesigns of the Internet. I argue that the Internet has no requirement for redesign based on its current set of uses. Applying security band-aids to the network will inevitably happen with or without our help, and is therefore a topic unworthy of the name "grand challenge". Conversely, redesigns of the Internet that are "cleaner" in some abstract way but do not enable new applications are poorly motivated.

However, understanding the planet in real time requires building a very different kind of network from the Internet. For one thing, computation is intimately embedded in the network - in-network processing of data is essential to correctly route the information, and reduce the data to enable the system to scale to the size of the planet.

Such in-network data processing implies that the dominant form of communication in such a network will be overlay-based (trees, meshes, hypercubes, etc.). Network protocols like TCP designed for 2-way, 2-node communication are inadequate, and research is needed into more appropriate ways to transmit data while sharing network resources in a useful way.

Security: Who will control this information? In this vision protecting ownership and privacy of information is arguably less important than authenticity or fidelity - lots of players have vested interests in influencing this data and its interpretation to their advantage.

In summary, such a challenge requires deep research in a variety of sub-fields of distributed systems: resource management, security, query processing, routing, data integration, discovery, naming, etc. etc. It also involves building a substantial artifact - network of processing elements which is connected to today's Internet, and can probably be bootstrapped from it. Sure it's plumbing, but plumbing is what we're good at, and this is plumbing on a grand scale.

I believe it is naive to think that we can improve the world by throwing more distributed systems at it, but distributed systems are the hammer we have, and this is the best-looking nail I can see. I'm also convinced that access to much more information by itself will not help the world, without a corresponding change in culture and thinking at a global scale. The optimistic view is that a project like this helps to facilitate such a change.