MareNostrum Supercomputer (a.k.a. Spain's Brain)
Putting the 'humanity' into HPC

Some exciting things are happening in Barcelona, Spain. In order of presumed interest to LinuxWorld readers, these are:

MareNostrum, the fifth fastest supercomputer in the world, which runs SuSE Linux on over 2,000 IBM blades, was recently christened in Barcelona.

Barcelona is undertaking a massive (approximately 1 square mile) urban revitalization program of an all-but-abandoned textile manufacturing district right in the center of the city, transforming it into a world-class Technology Park (www.bcn.es/22@bcn).

The Spanish government is moving their equivalent of the Federal Communications Commission from Madrid, the nation's capitol, to Barcelona. Given the historic rivalry between these two cities, this move packs the emotional equivalence of the New York Yankees moving to Boston.

Taken together, these initiatives position Barcelona at the vanguard of an economic renaissance built on the belief that the prosperity and well being of future generations will be attained through the advancement of knowledge. The crown jewel in this strategy is clearly the MareNostrum supercomputer.

MareNostrum, which means "our sea" in Latin (referring to the Mediterranean), is run by the Barcelona Supercomputing Center (www.BSC.es), a joint venture of the Spanish central government, the regional government of Catalonia, and the Polytechnic University of Catalonia, or UPC. Constructed out of 2,400 IBM JS20 blade servers running SuSE 9 Linux, MareNostrum went from design to "On" in less than 8 months.

Projects under way at BSC are pushing back the clouds of ignorance in areas as diverse as bioinformation, earth sciences, and fluid dynamics. And requests are pouring in from around Spain and Europe to leverage MareNostrum's mammoth smarts to unlock mysteries as diverse as which genes are behind the world's deadliest diseases to determining with greater certainty the potential impacts of, and contributors to, global warming.

Not Your Father's Supercomputer

The true significance of MareNostrum, however, goes well beyond its stature as the fastest computer in Europe, though this is in and of itself newsworthy. It goes beyond even the life-changing impact that the research done with its massive 27.91 Teraflops (trillion floating-point calculations per second) of processing power will have on the people of Spain, Europe, and the World, though this, too, is material for several articles. The true significance of MareNostrum lies in the fact that it heralds a new era in the accessibility of massive computing power to provide answers to the world's most vexing questions. This accessibility of supercomputing power marks a watershed event, an "inflection point," as the BSC's Felipe Lozano put it, in the impact that supercomputing power will have on society.

MareNostrum is situated in the public domain. Contrast this with the Earth Simulator, the world's fourth fastest supercomputer. Its name alone points to the most profound difference between MareNostrum and it, and most other top-ranked supercomputers before MareNostrum: that is, the majority of previous systems have been purpose-built to advance science related to a specific area of study. This is true of NASA's Columbia, the third fastest, and of the U.S. Department of Energy's Thunder, the seventh fastest. (Notable exceptions to this rule are the several university-based supercomputers in the Top 500, such as Virgina Tech's Dual 2.3GHz Apple Xserve system (ranked 14th on the Top 500), the San Diego

Supercomputer Center (43rd). These centers typically are available to academic researchers either on a per-use fee basis and/or whose projects are funded by grants and awards.)

Fasten Seatbelts

Think of application-specific supercomputers as the equivalent of air travel before the arrival of airlines, before air routes, before air traffic control, and before airports. Back then, air travel was reserved for the elite few who could afford not only an airplane, but a staff to maintain and fly it, facilities to store it, and purpose-built space on which to land and takeoff. Were planes back then fast relative to the other modes of transportation? You bet. Were they a quantum leap forward in transportation? Yes. But aviation technology did not, indeed could not, have the massive impact on society that it has had until air travel became accessible to more of society. And this did not occur until three key things happened, which together resulted in an air travel boom:

- ***Supply:*** The availability of comfortable, reliable, large, fast, safe, and efficient planes increased air system capacity and brought airfares down.
- ***Demand:*** Passenger interest in air travel rose as people began to see applications for this transportation mode for their own business and personal lives, and lower prices made it affordable.
- ***A little help from Uncle Sam:*** Governments reduced the cost and risk of running an airline by using taxpayer dollars to build airports and by regulating market entry.

In this context, MareNostrum is part of a revolution in supercomputing equivalent to the arrival of airlines, to the supply of airplanes built for transporting passengers, and to the decision by governments to ensure that air travel would, ahem, take off. Each of the three factors that contributed to greater accessibility of air travel, and to its subsequent boom, can be seen in the MareNostrum project.

Supply

Just as advances in aeronautical science allowed larger planes to carry more people faster while consuming less fuel, so too is computational science increasing the number of processes per second while simultaneously reducing footprint and power consumption. The computers that comprise MareNostrum, IBM's BladeCenter JS20, are certainly advanced. With two 2.2 Gigahertz (GHz) PowerPC 970 FX chips on each hot-swappable blade, they feature integrated 2-port Gigabit Ethernet, 4GB of memory per node, Fiber Channel expansion capability, and a host of reliability and management features. What is equally impressive about this technology is that it is available for any business to purchase. Having been designed for commercial use means that the JS20, like other blade solutions on the market today, is reliable, efficient, and compact: all requirements of the business computing market.

For the builders of MareNostrum, this meant that, unlike other supercomputers that must be housed in large, custom, and expensive facilities (just the building that houses DoE's Thunder cost over $50 Million), MareNostrum faced far fewer restrictions. This is a key contributor to the accessibility factor: without the need for elaborate and expensive facilities, a major barrier to implementing supercomputing power is torn down. Thus freed from these facilities shackles, the Catalan's aesthetic flair was on full display with their selection of a renovated cathedral to be MareNostrum's home.

Computational Performance - It's the System, Stupid...

As important as the commercial availability of efficient and powerful systems was to MareNostrum's success, advances in coordinating the massive quantities of processing power were equally important. MareNostrum utilizes a Myrinet Fiber Channel grid communication system that allows each blade's processors to communicate with any other blade's processors at a rate of four gigabits per second. To ensure that all the system's processors work together, tasks must be divided into 4,800 equal parts and each processor then writes its results to a shared hard drive memory system. (Smaller tasks that do not require the entire MareNostrum system are divided into equal parts equivalent to the number of processors they will be using. In this way, MareNostrum can work simultaneously on several different projects.)

As the BSC's Executive Director, Mateo Valero, put it in an interview with Spain's El Pais newspaper, "There is no other supercomputer in the world with the quantity of shared processors as MareNostrum. This model changes the philosophy of computation." One way it changes the philosophy of computation is that no longer is processor speed the limiting, or indeed the driving, factor to computational power. With the ability to divide a workload and distribute it equally across thousands of processors, all of which write their results to a common hard drive system, MareNostrum is one of a growing set of examples of how Moore's Law may be reaching the end of its usefulness, at least as it pertains to supercomputing. Dave Turek, IBM's vice president of Deep Computing, notes that IBM's Blue Gene systems, which took the top two spots in the most recent Supercomputer Top 500 listing, epitomize this point. These systems run miserly 700 MHz PowerPC chips.

Returning to our air system analogy for a moment, a Moore's Law equivalent here would seem to be to measure the advancement of aeronautical science solely by increases in jet engine horsepower. Clearly, jet engines have become more powerful over time, and clearly this has contributed significantly to the growth of the air system. But the ability of airlines to be commercially viable while carrying ever-greater numbers of passengers required advances system-wide. More powerful jets were needed, but they were no more important an ingredient in the booming capacity of the air system than were stronger and lighter aircraft materials, more aerodynamic plane designs, and better air traffic control technology to coordinate the larger number of planes in the sky.

MareNostrum achieves its immense 27-plus Teraflops using relatively slow 2.2GHz processors - by comparison, Intel's fastest chip does 3.66GHz, 66% faster than the chips in MareNostrum. As Mr. Turek explains, "to be a viable solution, MareNostrum had to be balanced - that is, it had to be powerful, compact, and energy efficient, and heat is the biggest enemy of energy efficiency. The faster the processor, the more heat it produces, which ruled out platforms based on faster, but yet much hotter, chipsets. The MareNostrum system requires an order of magnitude less power than the next fastest Supercomputer, which translates into millions of dollars in ongoing operational savings for BSC."

Demand

With the cost of supercomputing reduced to the time and thought it takes to submit a proposal to the BSC Selection Committee, there is no shortage of demand for access to MareNostrum. Still, it may take time for the full import of access to MareNostrum to sink in to the Spanish and European research community. The first set of projects will likely be evolutionary in nature - that is, they will use MareNostrum to accelerate existing lines of research. In time, though, the prospect of accessing MareNostrum will surely cause scientists to rethink what's possible, yielding a more revolutionary wave of research topics.

Government Support to Share Risk

Governments have played a major role in many of the most historically significant breakthroughs in technology that spurred innovation and ultimately had a profound impact on society. Dr. Rob Atkinson, head of the Progressive Policy Institute's Technology and New Economy project, documents this in his new book The Past and Future of America's Economy: Waves of Innovation that Power Cycles of Growth (www.ppionline.org/ppi_ci.cfm?knlgAreaID=107&subsecID=123&contentID=253184):

The telegraph got a major boost when Samuel Morse received a government grant to build a 40-mile long telegraph line between Washington and Baltimore in 1843. Herman Hollerith, a former employee of the Census Bureau, developed an automated counting machine using punch cards for the 1890 census, which saved the government $5 million. He later sold his company to a conglomerate that became IBM. The Wright Brothers owed the continued development of the airplane to Department of Navy contracts. A defense-sponsored microwave research program at Columbia University uncovered the basic concepts that led to the laser...Database technology owes its start to government funding. In fact, the leading database company, Oracle, grew out of a Central Intelligence Agency-funded database *software project in the 1970s. The World Wide Web was developed by Tim Berners-Lee at CERN, a European government-funded research institute. Even Google, the popular search engine company, owes its origins to knowledge that was developed with federal funds.*

Government support for these innovations was critical because, in all of the cases, the innovation was risky, the benefits were widely diffused beyond what an individual company could capture, and the innovations had infrastructural characteristics that laid the ground for innovation in a host of other industries and applications. MareNostrum has all three characteristics.

Jerry Sheehan, principal administrator/senior economist in the OECD's Science and Technology Policy Division in Paris, adds that centers such as MareNostrum also play a key role in economic development in at least two ways. "First, they serve as a magnet for talent, which is especially important in Europe, where EU nationals can work in any EU country. Hence, a centre like MareNostrum can help Spain attract researchers from across Europe. Second, since Spain produces a good number of researchers each year in a variety of fields, this kind of centre can provide them with much needed jobs. This can offset the brain drain that Spain has experienced in the past as its skilled researchers, trained largely at taxpayer expense, headed elsewhere in Europe or to the U.S. to find jobs."

The Acid Test: Utilization

As with air travel, where the true measure of its impact was neither the speed nor the volume of the planes, but rather the total number of passengers carried, so, too should we measure the advancement of supercomputing not merely by the number of computing operations accomplished per second, but also by the total number of projects carried out. With supercomputing reaching a point where an increasing percentage of members of the Top 500 are, like MareNostrum, shared resources, the ratio of projects to supercomputers will increase dramatically.

Consider this: an application-specific supercomputer, such as Thunder and Columbia, will only ever carry out research in a limited area. Now think about the fact that there is a list of over 100 projects from diverse disciplines waiting to take advantage of MareNostrum, and the folks at BSC figure they can handle in the neighborhood of 25-50 projects per year. With an estimated useful life of five years, this means that MareNostrum - one single supercomputer - holds the promise of promoting up to 250 different lines of research. Fasten Seatbelts.