Changing problems, databases, and tools in spatial search

W. Randolph Franklin

Rensselaer Polytechnic Institute
Troy, NY, USA
• Philosophically a Computer Scientist.
• PhD officially in Applied Math.
• Working in Electrical, Computer, and Systems Engineering.
• Teaching Engineering Parallel Computing.
• Collaborating with Geographers for 45 years.
• Implemented the first Triangulated Irregular Network (TIN) in geography, in 1973.
• Enjoy applying computer science and engineering to GIS.
Important concerns in Spatial Search

New:

- problems.
- hardware.
- SW tools.
- algorithms and data structures.
New search problems

- Coincidences in temporal track logs:
  - past near collisions between aircraft — but note the distance metric.
  - \( \geq 5 \) people from a group of 1000 flocked together.
- Point inclusion against \( 10^6 \) polygons — if this is a function in a larger system, then unlikely errors cannot be overlooked.
- Possible synergy with Computer Aided Design etc.
Available hardware is changing

Why parallel HW?

- More processing $\rightarrow$ faster clock speed.
- Faster $\rightarrow$ more electrical power. Each bit flip (dis)charges a capacitor through a resistance.
- Faster $\rightarrow$ requires smaller features on chip
- Smaller $\rightarrow$ **greater** electrical resistance!
- $\implies \iff$

- Serial processors have hit a wall.

![Diagram of circuit with capacitor and resistance](image.png)
Parallel HW features

- IBM Blue Gene / Intel / NVidia GPU / other
- Most laptops have NVidia GPUs.
- Thousands of cores / CPUs / GPUs
- Lower clock speed 750MHz vs 3.4GHz
- Hierarchy of memory: small/fast → big/slow
- Communication cost ≫ computation cost
- Efficient for blocks of threads to execute SIMD.
- OS:
  - 1st through 186th fastest run variants of Linux.
Massive Shared Memory

- Massive shared memory is an underappreciated resource.
- External memory algorithms are not needed for most problems.
- Virtual memory is obsolete.
- An 80-core workstation with 1TB (1 million million bytes) of memory costs $40K.

Code

```c
const long long int n(5000000000);
static long long int a[n];
int main() {
  double s(0);
  for (long long int i=0; i<n; i++)
    a[i] = i;
  for (long long int i=0; i<n; i++)
    s += a[i];
  std::cout << "n=" << n << " , s=" << s << std::endl; }
```

Runtime: 60 secs w/o opt to loop and r/w 40GB. (6 nsec / iteration)
Larger databases to search

- LIDAR — $10^5 \times 10^5$ terrains.
- New York City taxi logs — 14 million trips in 2013.
- Streaming sensors — process it in real time or lose it.
Tools to Access the HW

- OpenMP
- CUDA
- Thrust

Projected Performance Development

(Highlights of the 42nd Top500 List, SC13)
const int n(500000000);
int a[n], b[n];
int k(0);

int main () {
    #pragma omp parallel for
    for(int i = 0; i < n; i++) a[i]=i;
    #pragma omp parallel for
    for(int i = 0; i < n; i++) {
        #pragma omp atomic capture (or critical)
        j = k++;
        b[j] = j; }
    double s(0.);
    #pragma omp parallel for reduction(+:s)
    for (int i=0;i<n;i++) s+=a[i];
    cout << "sum: " << s << endl; }
Large systems and packages on which to realize your implementations.

- **gmp++** – big rationals.
- **Computational Geometry Algorithms Library (CGAL).**
- **Matlab** — state-of-the-art numerical code for matrices.
- **Mathematica** — math, plots, etc, etc.

Commercial tools are expensive.
Simple regular data structures and algorithms

- Adapting to the HW
  - farm out work to the parallel processors
  - data movement expensive
  - many processors executing the same code.
- Keep it simple.
- Uniform grid good.
- Trees not so good.
Important concerns in Spatial Search

New:

• problems.
• hardware.
• SW tools.
• algorithms and data structures.