# High Performance Computing for Disease Surveillance

**David W. Bauer[1], Ph.D., Brandon W. Higgs[1], Ph.D., Mojdeh Mohtashemi[1, 2], Ph.D.**

[1]*MITRE;* [2]*MIT CS and AI Lab*

## OBJECTIVE

Space-time detection techniques often require computationally intense searching in both the time and space domains [1, 2]. We introduce a high performance computing technique for parallelizing a variation of space-time permutation scan statistic applied to real data of varying spatial resolutions and demonstrate the efficiency of the technique by comparing the parallelized performance under different spatial resolutions with that of serial computation.

## BACKGROUND

Space-time detection of disease clusters can be a computationally intensive task which defies the real time constraint for disease surveillance. At the same time, it has been shown that using exact patient locations, instead of their representative administrative regions, result in higher detection rates and accuracy while improving upon detection timeliness [3]. Using such higher spatial resolution data, however, further exacerbates the computational burden on real time surveillance. The critical need for real time processing and interpretation of data dictate highly responsive models that may be best achievable utilizing high performance computing platforms.

## METHODS

Traditional approaches towards parallelizing an application fall into two categories: (*i*) *data* and (*ii*) *algorithm decomposition*. Because of the high degree of space-time dependency within the scan statistic algorithms, here we focus on decomposing the application in the data domain. The data are provided by the San Francisco Department of Public Health (SFDPH), Tuberculosis Program. The geospatial information in the data consists of precise locations of 392 homeless individuals infected with tuberculosis (TB). We mapped each individual location to the corresponding census tract using ArcGIS v9.2 (ESRI). Both individual locations and census tracts were used as spatial bases for comparing the speedup and efficiency using parallel computation. Parallelization was performed on a Linux Beowulf cluster with a total of 64 processors. By Amdahl's law [4], if $F$ is the portion of an application that is sequential, then the maximum speedup $S_{\max}$, and efficiency $E_N$, then using $N$ processors:

$$S_{\max} = \frac{1}{F + \frac{1-F}{N}}, \quad E_N = \frac{S_{\max}}{N}$$

The total number of spatial scanners sampled for the census tracts was 441, while the total for individual addresses was 4,234. The time window varied from 4 to 72 weeks spanning a period of ten years. The detection algorithm was thus parallelized by processing each space-time pair (cylinder) onto a different CPU.

## RESULTS

Figure 1 illustrates that using 16 processors achieves an efficiency of 93% based on census tracts, and 100% using exact coordinates, resulting in a respective 15-fold and 16-fold improvement over sequential execution. Parallelization using individual addresses is more efficient due to higher computational fidelity. When we utilize 32 or more CPUs, the efficiency drops respectively to 45% and 70%, indicating that there is a large sequential component to the data set that cannot be further parallelized (not shown).

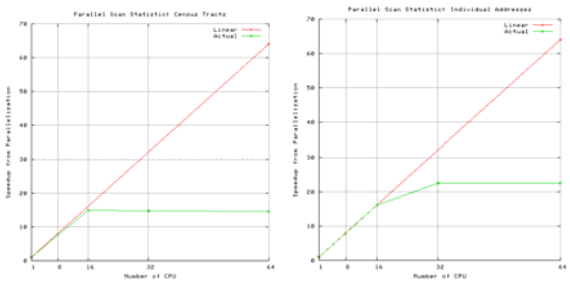

Figure 1 – Comparison of parallel performance based on census tract and individual address to serial.

Decomposing the data domain into individual addresses allows for a higher degree of parallelism within the application. Figure 1b shows an efficiency of 68.75% for 32 CPU, an improvement over census tracts, and a 22-fold improvement in the runtime.

## CONCLUSIONS

We proposed a highly efficient parallel computation technique for real time processing of a space-time outbreak detection technique. We have shown that high performance computing platforms, such as a Linux Beowulf Cluster, can come closer to meeting the needs of real time surveillance constraints. In the future we plan to investigate methods of parallelizing the scan statistic algorithm, to further increase the efficiency of the model on multiple processors.

## REFERENCES

[1] Weinstock MA.: A generalized scan statistic test for the detection of clusters. In International Journal of Epidemiology (1982) 10:289-293.

[2] Kulldorff M, Heffernan R, Hartmann J, Assuncao R, Mostashari F: A space-time permutation scan statistic for disease outbreak detection. In Public Library of Science (2005) 2(3).

[3] Tuberculosis Outbreaks among the San Francisco Homeless: Trade-offs Between Spatial Resolution and Temporal Scale (under review)

[4] Amdahl G. Validity of the single processor approach to achieving large-scale computing capabilities. In AFIPS Conference Proceedings (1967) 30:483-485.