Son muchas las áreas de investigación, y sobre todo, programas de aplicación real donde la capacidad de cómputo de un único procesador no es suficiente. Un ejemplo lo encontramos en el campo de la bioinformática y, en particular, en la rama de genómica, donde los secuenciadores de genomas son capaces de producir millones de secuencias en un día. Estos millones de secuencias se deben procesar y evaluar con tal de formar el genoma del ser vivo que se haya analizado. Este procesar y evaluar requieren un tiempo de cómputo y capacidades de memoria muy grandes, sólo al alcance de computadores con más de un core y una capacidad de almancenamiento significativa. Otros ámbitos donde nos podemos encontrar con estas necesidades de cómputo son en el diseño de fármacos, estudios del cosmo, detección de petróleo, simulaciones de aeronaves, etc.

Durante años, el aumento de la frecuencia de los procesadores había sido la vía de aumentar el rendimiento de las aplicaciones, de una forma transparente al programador. Sin embargo, aunque la frecuencia de los uniprocesadores ha ido aumentando, se ha observado que este aumento no se puede mantener indefinidamente. No podemos hacer mover los electrones y los protones a mayor velocidad que la de la luz. Por otra parte, este aumento de frecuencia conlleva un problema de disipación de calor, haciendo que los uniprocesadores deban incorporar mecanismos de refrigeración de última generación. Finalmente, la mejora de la tecnología y por consiguiente, la consecuente reducción del tamaño del transistor, nos permiten incorporar más componentes al uniprocesador sin aumentar su área (más memoria, más *pipelines*, etc.). Sin embargo, esta disminución del tamaño del transistor para incorporar más mejoras y capacidades al uniprocesador tampoco es una solución que se pueda llevar al infinito (principio de incerteza de Heisenberg). Estos problemas tecnológicos contribuyeron a que, para poder tratar con problemas tan grandes evitando las limitaciones tecnológicas, los arquitectos de computadores comenzaran a centrar sus esfuerzos en arquitecturas paralelas.

El soporte hardware para procesar en paralelo se puede introducir en varios niveles; desde el más bajo nivel, el uniprocesador, al más alto nivel con las grids de multiprocesadores o multicomputadores, donde la red de interconexión es Internet. Estos niveles abarcan el paralelismo a nivel de instrucción (ILP), a nivel de datos (DLP) y a nivel de threads (TLP).

Principio de incerteza de Heisenberg

"The position and momentum of a particle cannot be simultaneously measured with arbitrarily high precision. There is a minimum for the product of the uncertainties of these two measurements. There is likewise a minimum for the product of the uncertainties of the energy and time."

Niveles de paralelismo

ILP: Instruction Level Parallelism.

DLP: Data Level Parallelism.

TLP: Thread Level Parallelism