Multiplicación de matrices

Tonny Gonzales Villafuerte

March 2017

1 The simple three-nested-loop version of the matrix product

Normalmente, un algoritmo que se refiere a elementos individuales es reemplazado por uno que opera en sub-arrays de datos, que se llaman bloques en el campo de cálculo de matriz. Las operaciones sobre sub-arrays se pueden expresar de la forma habitual. La ventaja de este enfoque es que los pequeños bloques se pueden mover a la memoria local rápida y sus elementos pueden ser utilizados repetidamente.

```
for(i=0; i<m; i++)
    for(j=0; j<n; j++)
    {
            z[i][j] = 0;
            for(k=0; k<c; k++)
             z[i][j] += x[i][k] * y[k][j];
    }
```

El cálculo completo 2n operaciones aritméticas (contando adiciones y multiplicaciones por separado), pero produce y consume sólo 3n valores de datos. En conjunto, el cómputo exhibe admirable reutilización de datos. En general, sin embargo, una matriz entera no cabrá en una memoria local pequeña. El trabajo debe por lo tanto ser dividido en pequeños trozos de cálculo, cada uno de los cuales utiliza una pieza suficientemente pequeña de los datos. Obsérvese que para cada iteración del bucle exterior (es decir, para un valor dado de i) se realizan operaciones n2 y se hace referencia a los datos n, sin reutilización. Para los valores fijos de i y j, n de cálculo y n datos referidos también - de nuevo, no hay reutilización.

2 The blocked version with six nested loops

En primer lugar, tenga en cuenta que en este programa se realizan exactamente las mismas operaciones en los mismos datos; Incluso el error de redondeo es

idéntico. Sólo la secuencia en la que se realizan operaciones distintas es diferente del programa desbloqueado. Todavía hay reutilización en todo el programa de orden n. Pero si consideramos una iteración con i0 fijo, j0 y k0, vemos que se realizan operaciones de 2b3 (por los tres bucles internos) y se informan los datos de 3b2. Ahora podemos elegir b lo suficientemente pequeño para que estos datos 3b encajen en la memoria local y así lograr la reutilización b-fold. (Si esto no es suficiente - si b ¡B en otras palabras - entonces la máquina está mal diseñada y necesita más memoria local.) Dicho de otra manera, si necesitamos la reutilización B-fold, elegimos el tamaño del bloque b = B.

```
for( i1=0;i1<(nn/BlockSize);++i1)
{
    for(j1=0;j1<(nn/BlockSize);++j1)
    {
        for(k1=0;k1<(nn/BlockSize);++k1)
        {
            for(i=i1;i<min(i1+BlockSize-1,nn);++i)
            {
                for(j=j1;j<min(j1+BlockSize-1,nn);++j)
                {
                    for(k=k1;k<min(k1+BlockSize-1,nn);++k)
                    {
                     cc[i][j] = cc[i][j] + a[i][k] * b[k][j];

                    }
                }
            }
        }
    }
}
```