ISCAS 中国科学院软件研究所学术年会'2023
暨计算机科学国家重点实验室开放周

学术论文

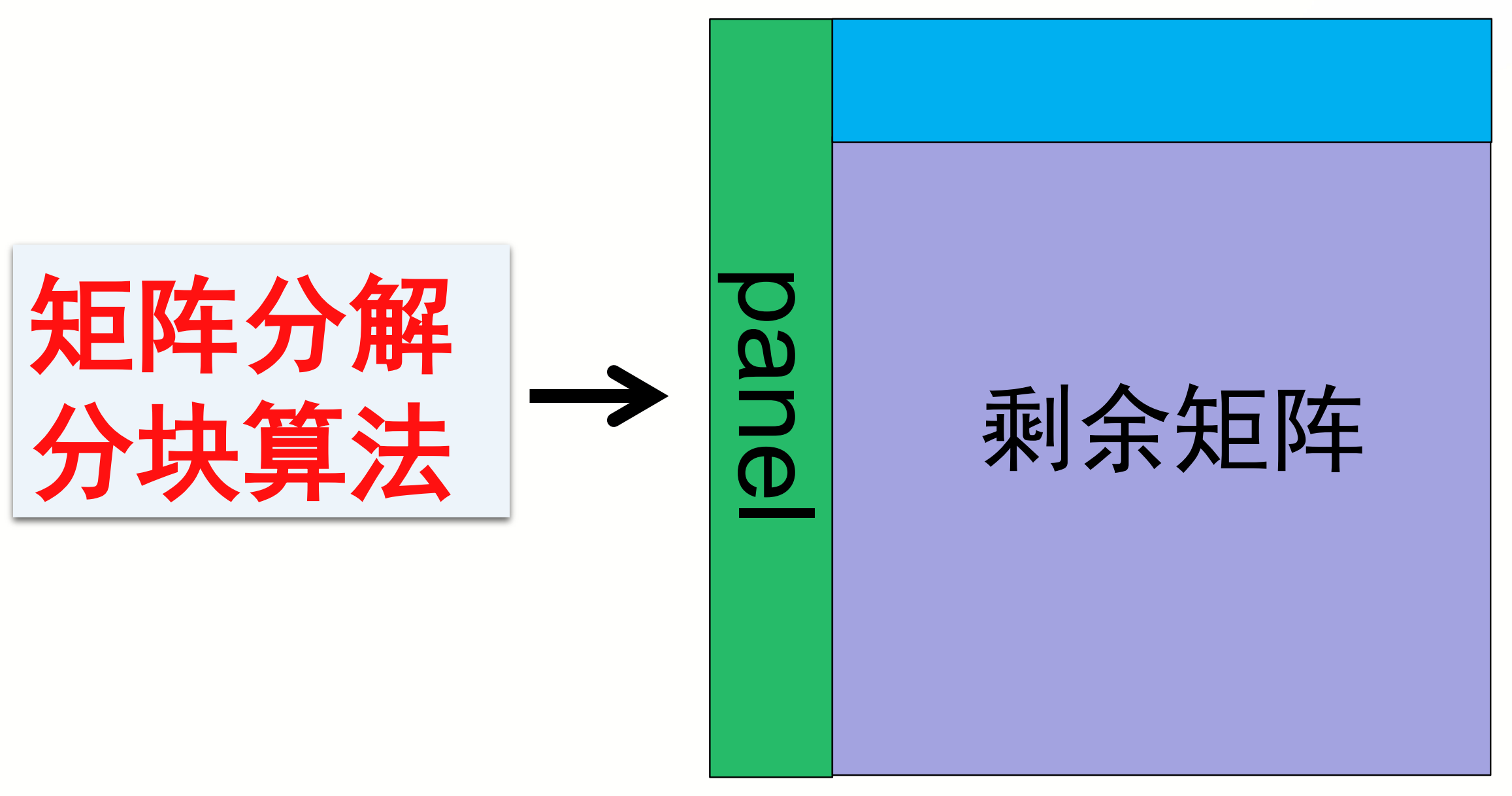
An Optimized Framework for Matrix Factorization on the New Sunway Many-core Platform
SW26010Pro众核处理器上矩阵分解函数优化框架

马文静，刘芳芳*，陈道琨，路青霖，胡怡，王鸿森，袁欣辉

ACM Transactions on Architecture and Code Optimization, Vol. 20, No. 2, Article 23. Publication date: March 2023.

联系人：刘芳芳，13466713051，fangfang@iscas.ac.cn

继神威太湖之光的申威26010处理器之后，新一代申威众核处理器——申威26010Pro闪亮登场。在这种独特的处理器上编写复杂程序有很大挑战，一个典型例子就是矩阵分解。我们根据矩阵分解在科学计算、人工智能等实际应用中的需求，设计了一套优化框架，可以方便地实现申威26010Pro上高效的矩阵分解代码。



剩余矩阵更新：对大块矩阵的操作，浮点计算占优，性能高

panel分解：矩阵-向量操作为主，访存占优，性能低

```
panel分解原始实现方式(LU分解)
for each column {
   //idamax和ger都是BLAS函数
   idamax();
   ......
   dger();
}
```

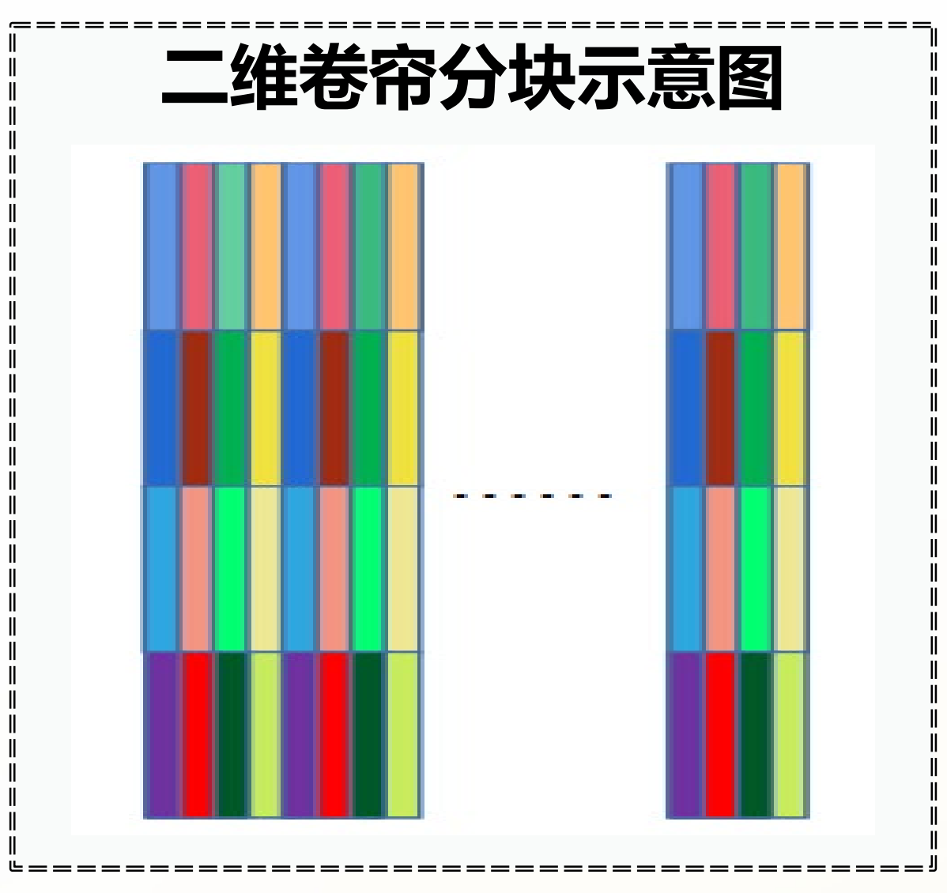
```
从核BLAS函数代码实现：
DMA_load();
//各从核完成计算任务
DMA_store();
```

大量DMA数据传输开销！

panel分解时间在中小规模计算占比尤其大

中小规模panel可放入64个从核的LDM中

将panel按二维卷帘分块分配到从核阵列。读入从核LDM后，从核内完成panel分解的全部迭代



panel分解所需的BLAS函数调用序列

提供从核内panel分解模板类，包含内置核间通信的参数化BLAS函数接口

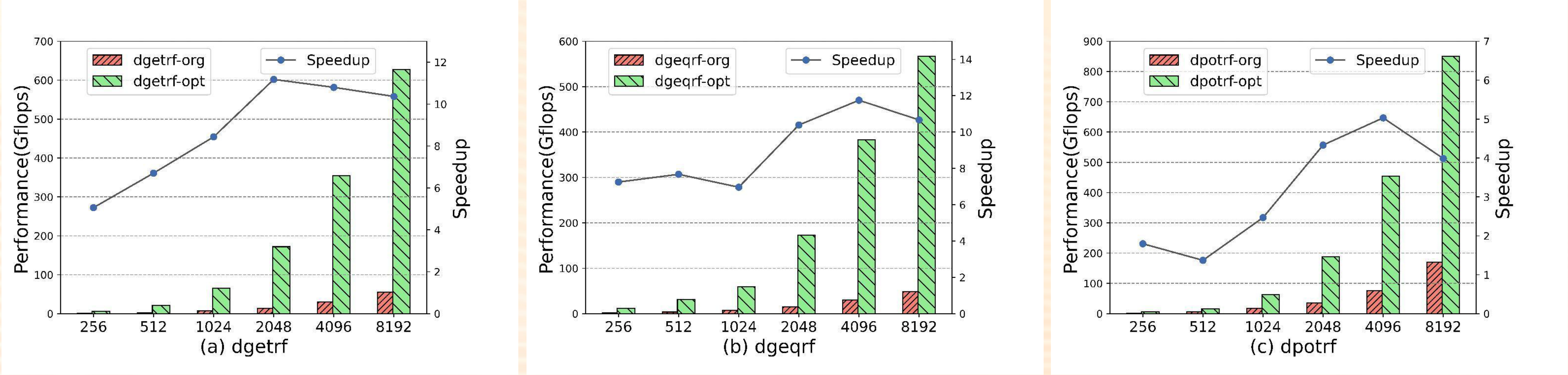
```
//使用模板及接口实现的从核内panel分解（LU分解）
void panel_factorization{
   DMA_load();
   for each column {
       if(_COL==cur_col)
           columm_bottom_func(idamax......)
           ......
       bottom_right_func(ger......);
   }
   DMA_store();
}
```

递归panel分解：保证矩阵更新操作的规模，并使用更窄的panel，以使更大的矩阵也能装入CPE的LDM中

改变矩阵更新算法实现方式：提高总体性能，为递归方案提供方便

单从核向量化BLAS计算代码：提高panel分解计算效率

对不同规模矩阵进行LU分解(dgetrf)、QR分解(dgeqrf)、Cholesky分解(dpotrf)，相比原始实现方式平均加速比分别达到9. 76，10. 12和4. 16



(a) dgetrf (b) dgeqrf (c) dpotrf