软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(Suppl.(2)):182−191

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

支持复杂负载模型的互连网络性能模拟研究*

胡 凯+, 陈陆佳, 王 哲, 蒋 树

(北京航空航天大学 计算机学院,北京 100191)

Study on Performance Simulation of Interconnection Network under Complex Workload Mode

HU Kai+, CHEN Lu-Jia, WANG Zhe, JIANG Shu

(School of Computer Science and Engineering, BeiHang University, Beijing 100191, China)

+ Corresponding author: E-mail: hukai@buaa.edu.cn

Hu K, Chen LJ, Wang Z, Jiang S. Study on performance simulation of interconnection network under complex workload mode. *Journal of Software*, 2011,22(Suppl.(2)):182−191. http://www.jos.org.cn/1000-9825/11039.htm

Abstract: The interconnection network's design of large-scale parallel computer systems is significant to the execturion of the efficiency of parallel programs. Currently, the Petaflop supercomputers usually have more than ten thousand computing nodes, which cause new challenges to the performance of interconnection network. However, most exited studies only consider simple network workload models which have many differences from the workloads of real parallel applications. This paper presents complex workload models which are more similar to the practical network workload. Then based on the mathematical model of the interconnection network in the earlier study, the study uses a flit-level network simulator to support analyzing those complex workload models. Finally, through a great deal of experiments the performance of Torus and Fat Tree network topologies are compared under different workload models. Meanwhile, the message mean latency of the 3D FFT parallel algorithm using 2D partitioning is simulated. The results and related analysis efficiently support the large-scale interconnection network's design, as well as the optimization of parallel programs.

Key words: interconnection network; performance analysis; network simulation; complex workload model

摘 要: 大规模并行计算机系统互连网络的设计对并行程序执行效率有重要影响,当前千万亿次计算机系统拥有上万个节点,给互连网络的性能带来新的挑战.然而,目前互连网络性能研究大多考虑消息目的地均匀分布等简单负载模型,这与真实并行应用的网络负载存在较大差异.首先,在简单负载模型的基础上,增加考虑局部通信、热点通信等因素,研究更接近真实网络负载特性的复杂负载模型.其次,基于此前对互连网络建立的数学模型,扩展了一个微片级网络模拟器,以支持模拟复杂负载模型;最后,通过大量模拟实验比较了 Torus 和 Fat Tree 网络在各种复杂负载模型下的性能,并以 3D FFT 算法的 2D 任务分解为例,模拟互连网络的消息延迟.模拟实验的数据和分析,给大规模并行计算机互连网络的设计、提升并行程序执行效率提供了帮助.

关键词: 互连网络;性能分析;网络模拟;复杂负载模型

* 基金项目: 国家自然科学基金(61073013); 航空科学基金(2010ZA04001)
收稿时间: 2011-07-15; 定稿时间: 2011-12-02

大规模并行计算机互连网络的设计对并行程序的执行效率有重要影响[1].当前,重要应用驱动并行计算机系统节点规模不断增长,例如IBM Blue Gene和Cray XT系列中的计算节点数目上万,节点间通信距离常多达几十跳,使得平均消息延时明显增加.因此,研究大规模并行计算机互连网络的性能,优化并行程序的通信有着重要的实际意义.

网络负载又对互连网络性能有很大影响.例如,一般情况下,对于给定目的节点分布,网络负载对虫洞交换网络平均消息延时的影响比其他设计参数的影响大得多,并且吞吐率主要受通信模式(目的节点的分布)的影响[2].因此,模拟网络负载对研究互连网络性能非常重要.目前已有的互连网络性能研究大多考虑目的节点均匀分布等简单负载模型,然而,大多数计算都呈现出某种程度的通信局部性.Johnson 研究了局部通信行为[3];Kim和 Chien 研究了网络中同时具有长、短两种消息情况时的性能[4];Loucif 等人研究了 Torus 网络中热点通信对消息延迟的影响[5].但是已有研究对更贴近真实并行应用的复杂负载模型刻画并不统一,也没有较全面的结果分析.本文基于此前对互连网络建立的数学模型[6],通过网络模拟的方法,研究复杂负载模型下互连网络的性能.

1 复杂负载下的互连网络性能研究

1.1 互连网络性能的影响因素

并行计算机互连网络的性能由其拓扑结构、路由算法、交换策略和流控机制共同刻画,它们的设计从不同侧面对网络性能产生重要影响[7].

拓扑结构是网络的物理互连结构,它可以是规则的,也可以是非规则的.目前大多数并行计算机采用高度规则的网络拓扑,例如 Torus,Fat Tree.路由算法决定消息在网络从源节点发送到目的节点通过的路径.交换策略决定消息中的数据如何穿越网络中的各条通道,常用的有存储转发、电路交换、虚切入和虫洞交换等.流控机制决定网络资源(通道和缓冲区)的分配原则,并对多个消息请求同一资源进行仲裁.虚通道技术将物理通道分时共享,是当前并行计算机网络中广泛使用的一种流控机制[8].

互连网络的性能还与并行计算机系统中具体运行的并行应用程序密切相关,不同的并行计算任务可能具有不同的通信模式,因此,评价互连网络的性能需要考虑能够代表典型并行应用通信特点的网络负载模型.

1.2 互连网络复杂负载模型研究

互连网络负载模型主要有 3 个参数:消息目的地分布、消息注入速率和消息长度分布[2].目前在研究互连网络性能时,大多考虑简单的负载模型,例如均匀的消息目的地分布、单一的消息注入模式以及固定的消息长度等,这与真实并行应用的网络负载存在较大差异.

消息目的地分布指明每个节点上,下一个消息传输的目的地.在研究互连网络性能时,均匀分布是最常用的分布.在这种分布下,对所有的 i 和 $j(i\neq j)$,节点 i 向节点 j 发送消息的可能性是相同的,而大多数计算都呈现出某种程度的通信局部性.互连网络节点间通信的局部性可分为空间局部性和时间局部性.当平均节点间距小于均匀分布时,呈现出空间局部性.当应用只与某个节点子集密切通信时,呈现出时间局部性.

在网络性能模拟中,所有节点的注入速率通常都是一样的.除了注入速率的大小,还应考虑源节点内相邻两个消息注入的时间间隔分布,它刻画了消息的注入模式,例如 Bernoulli 过程、Poisson 过程等.Bernoulli 过程中假设源节点每个周期最多只能产生一个消息,这种假设对真实的网络系统可能并不成立,更多的互连网络性能研究常假设消息产生过程满足 Poisson 过程.

消息长度分布是由网络设计中的许多要素和具体的并行应用共同决定的,在大多数模拟器中,消息长度的选择是固定的.然而研究长度差别很大的消息之间的相互影响也很有意义.例如,小部分极长消息可能大大增加了短消息的延时.有研究表明,长、短两种消息混合分布更接近于并行计算机网络的真实情况[4].

本文研究的复杂负载模型是在之前研究中常讨论的简单负载模型的基础上,增加考虑多种消息注入模式混合、局部通信、热点通信、长短消息混合分布等因素的负载模型.

2 互连网络性能模拟研究

2.1 互连网络性能的研究方法

研究并行计算机互连网络性能的主要方法包括网络测量(measurement)、分析模型(analytical model)和网络模拟(simulation)等.

网络测量需要在真实系统部署完成后通过专门的仪器设备进行测量,研究网络性能的准确性最高.分析模型通过建立数学模型分析网络性能,虽然方便设定参数,时间开销最少,不存在可扩展性的问题,而且能够利用数学理论给出系统设计的各个要素对网络性能的影响关系,但它相对网络模拟具有准确性差的缺点[9].

基于软件的网络性能模拟因为灵活、高效等优点,成为主流研究比较常用的方法.网络模拟与分析模型一样,广泛使用在并行计算机网络的设计阶段,也常用于验证分析模型计算结果的正确性.相对于分析模型,网络模拟花费的时间较长,但对真实系统性能的预测具有较高的准确性.因此,考虑将网络模拟和分析模型相结合的方法进行研究.在之前的研究中,我们针对 Torus 网络局部通信建立了数学模型,研究其对消息延迟和网络最大吞吐量的影响规律[6],通过数学模型对网络性能进行分析,本文主要采取网络模拟的方法研究互连网络性能.

2.2 网络模拟器的种类

网络模拟是一个涉及理论与实现的重要研究方向,使用网络模拟器是网络模拟的重要手段.文献[10]将常用的网络模拟软件分为 4 类:基于通用编程语言开发的模拟软件;使用专用模拟语言;支持模拟的软件库;模拟的软件包.从模拟的层次上分,网络模拟器可以分为包级(packet-level)、微片级(flit-level)和门级(gate-level)[11].对于包级网络模拟器,网络中传输的最小单位是消息包或整个消息,例如 BigSim 模拟器.门级模拟器模拟网络的层次最低,正确性最高,常用于设计和完善并行计算机互连网络的交换机.常见的门级模拟软件有 Orion 等.

微片级网络模拟器介于包级模拟器和门级模拟器之间.已有不少微片级模拟器被实现用于研究,例如 pp-mess-sim,SWORDFISH,SMART,PEPE,Pertel,SimRed,RSIM,Booksim 等.Booksim[7]是 Stanford 大学 Dally 等人开发的微片级并行计算机互连网络模拟器.SimRed[11]是西班牙 Valencia 大学 Duato 研究组近年来开发的微片级网络模拟器,提供图形界面和分别用 C++和 Java 编写的源代码.但是该软件的代码可读性不及 Booksim,支持模拟的选项也略少于 Booksim.

2.3 复杂负载下的互连网络性能模拟实现

基于 Booksim 具有良好可读性和容易扩展功能的特点,本文选择对 Booksim 进行扩展设计,增加部分功能和参数,以支持复杂负载的互连网络性能模拟,主要表现在以下几方面:

(1) 支持模拟 Poisson 过程的消息注入模式

消息注入速率是网络负载模型的 3 个要素之一,注入速率越大表示网络负载越重.通过模拟软件有两种常见消息注入模式:Bernoulli 过程和 Poisson 过程.Booksim 只支持消息以 Bernoulli 过程和 on & off 过程产生,因此增加了支持消息以 Poisson 过程产生的功能.

(2) 支持模拟长、短两种消息长度混合分布

Booksim 只支持网络单一长度的消息模式,即每一次模拟过程中节点只能产生一种长度的消息,因此增加功能,使模拟器支持模拟网络中同时存在两种长度消息的模式,并且可在配置文件中通过参数分别设置第 1 种、第 2 种消息长度和第 1 种消息的比例.

(3) 支持模拟局部通信及热点通信

Booksim 支持多种消息目的地分布,但是并未考虑局部通信和热点通信,因此实现了分别支持模拟局部通信和热点通信的功能.图 1 是产生局部通信消息函数的伪代码.

(4) 支持模拟 3D FFT 的 2D 任务分解

为结合实际应用背景研究网络性能,实现模拟 3D FFT 采用 2D 任务分解时两阶段通信模式的功能.

(5) 支持模拟两个特定节点间消息延迟

Booksim 主要研究网络宏观上的消息延迟,并不十分关心某两个节点间消息延迟的性质,但是实际研究可能需要了解两个特定节点间消息延迟的情况,而模拟过程并不能保证一定产生从特定源节点到特定目的节点的消息.于是加入模拟两个指定节点间消息延迟的机制,即在原始模拟过程中仅多产生一个或几个从指定源节点到指定目的节点的消息,这种机制对网络本身负载特性的影响很小.

```
number=网络的节点总数;
if 该消息是局部消息 {
    for( i=1; i < number; i++) {
        do 计算i到源节点的距离dist[i];
        if  dist[i]在局部通信域内
            do 记录i在局部通信域local_domain内;
    }
    在local_domain内,按目的均匀分布产生消息;
}
else
    在整个网络中,按目的均匀分布产生消息;
```

Fig.1 The method to generate messages of local communication

图 1 产生局部通信消息函数的伪代码

3 互连网络性能模拟结果分析

目前,Torus 和 Fat Tree 是大规模并行计算机互连网络设计中常用的拓扑结构,例如 Blue Gene 和 Cray XT 系列采用 3D Torus 或 3D Mesh,Roadrunner 和曙光 5000A 则采用 Fat Tree 结构.本文以 Torus 和 Fat Tree 网络为例,通过大量模拟实验比较两种网络在各种复杂负载模型下的性能.

3.1 目的地均匀分布模式下的互连网络性能

3.1.1 多种消息注入速率

图 2 是 4 096 节点 Torus 和 Fat Tree 网络不同消息注入模式下的负载-延迟曲线.实验结果表明,当网络规模较大、消息较长时,Bernoulli 过程下消息的延迟显著低于 Poisson 过程,同时支持更高的网络吞吐量.另外,均匀负载模式下,Torus 网络和 Fat Tree 网络的消息平均延迟存在一定差异,对于大规模网络,Fat Tree 的性能往往更优.

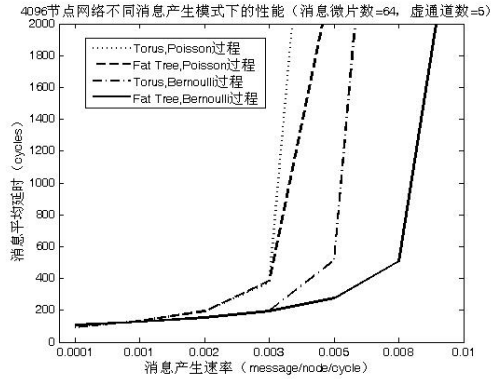
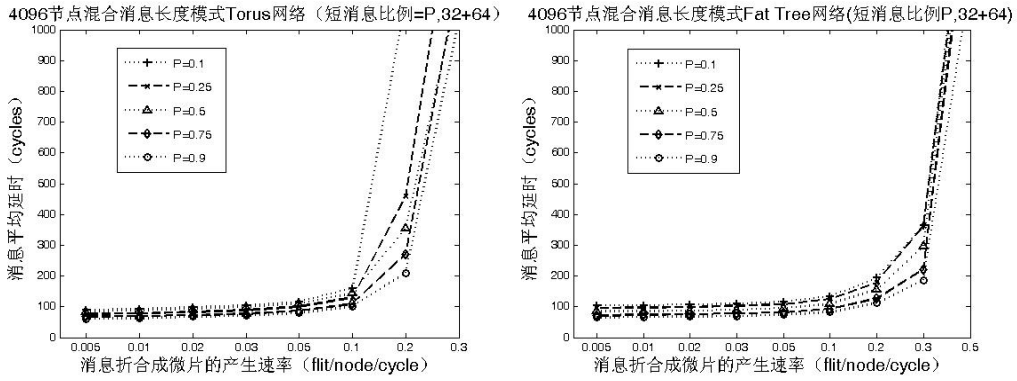


Fig.2 Latency curves under different message generated patterns (4 096 nodes)

图 2 4 096 节点网络不同消息产生模式下的性能

3.1.2 消息长度混合分布

图 3 是 4 096 节点 Torus 和 Fat Tree 网络不同消息长度分布下的负载-延迟曲线.实验结果表明,在相同网络负载下(源节点产生消息微片速率相同),随着短消息比例的增加,Torus 和 Fat Tree 网络的性能都得到改善,其中 Torus 相比 Fat Tree,对参数的变化更敏感.



(a) Torus (b) Fat Tree

Fig.3 Latency curves under mixed different length of message

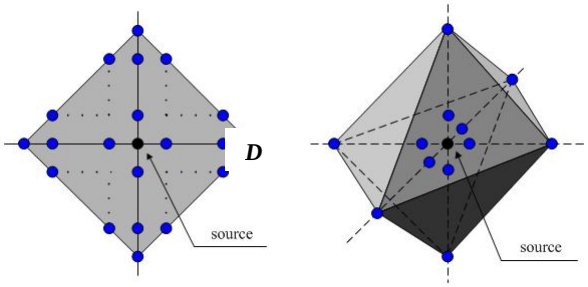
图3 网络混合消息长度模式下的性能

3.2 目的地非均匀分布下的互连网络性能

上节中均假设消息满足目的地分布均匀,即源节点产生的消息以等概率选择网络中的全部节点作为目的地.然而真实并行应用中节点间的通信行为往往具有局部性,例如微分方程的求解.研究非均匀负载模式下互连网络的性能具有实际的意义.本文重点研究了两类典型的非均匀负载模式:局部通信和热点通信,以及它们对网络性能的影响规律.

3.2.1 Torus 网络局部通信模式下的性能

定义二元参数 (D_{local}, R_{local}) 刻画 Torus 网络中局部通信的强度.假设源节点以固定概率随机产生两类消息:普通消息和局部消息,其中局部消息的比率 R_{local} 称为局部通信率.局部消息以等概率选取以源节点为中心,半径为 D_{local} 的球形区域内节点作为目的节点,称该区域为局部通信区域, D_{local} 为局部通信区域半径.图 4 分别为 2 维和 3 维 Torus 网络局部通信区域示意图.



(a) 2D Torus (b) 3D Torus

Fig.4 Illustration of local communication domain

图4 局部通信区域示意图

图 5 为 512 节点 Torus 网络局部通信下的性能曲线.实验结果表明,并行应用的局部通信能够有效利用网络带宽,缩短消息延迟,提高网络最大吞吐量;对 Torus 网络,二元参数中的局部通信率比局部通信区域半径对消息延迟的影响更明显,这表明,在优化并行应用的通信时,应首先提高具有空间局部特性的消息比例.

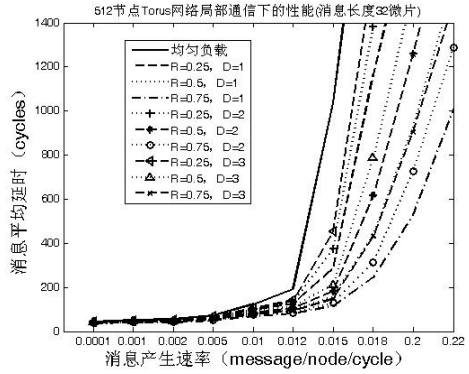


Fig.5 Latency curves under communication locality in Torus network

图5 Torus网络局部通信模式下的负载-延迟曲线

3.2.2 Fat Tree 网络局部通信模式下的性能

与 Torus 网络类似,使用二元参数 (D_{local}, R_{local}) 刻画 Fat Tree 网络中局部通信的强度. R_{local} 为局部通信率.局部消息以等概率选取源节点所在的高度为 D_{local} 层的 Fat Tree 子树内节点作为目的节点.称该区域为局部通信区域, D_{local} 为局部通信区域层数.图 6 为 Fat Tree 网络局部通信区域示意图.

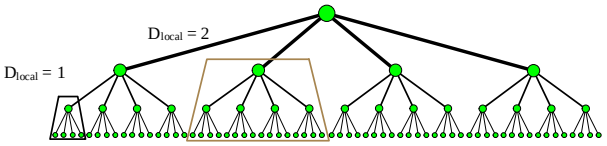


Fig.6 Illustration of local communication domain in Fat Tree network

图6 Fat Tree局部通信区域示意图

图 7 为 512 节点 Fat Tree 网络局部通信下的性能曲线.实验结果表明,与 Torus 网络类似,Fat Tree 网络中局部通信能够有效提高网络性能;但与 Torus 网络相反的是:在 Fat Tree 网络中,局部通信区域层数对消息延迟的影响更明显,因此应重点优化并行任务映射策略,以缩小 D_{local} 的数值.

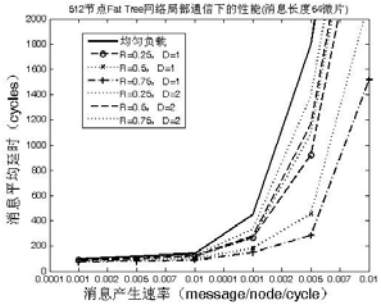


Fig.7 Latency curves under communication locality in Fat Tree network

图7 Fat Tree网络局部通信模式下的负载-延迟曲线

上述研究为具有局部通信性质的并行应用,给出了一种有效预测消息延迟和网络最大吞吐量的方法.

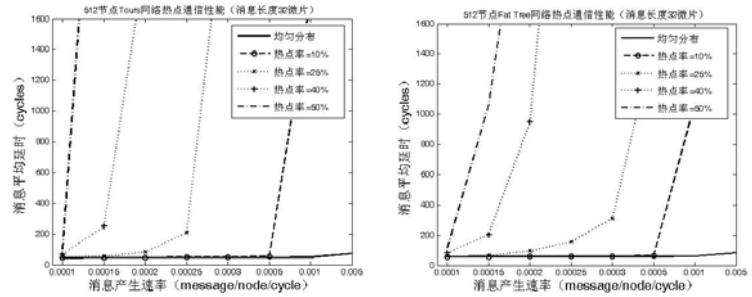
3.2.3 热点通信

热点通信是另一种并行应用中广泛出现的通信模式,例如 Barrier 同步的前半部分就可以看成是一个典型的热点通信过程.这里模拟了网络中存在一个通信热点时的性能,引入参数热点率刻画热点通信的强度:

热点率=每个源节点产生以热点为目的地的消息占其产生全部消息的比例.

考虑 512 节点 Torus 和 Fat Tree 网络,实验的性能曲线如图 8 所示.根据模拟结果,热点通信会大大增加消息

平均延迟,并导致网络在较低负载就达到饱和,尤其是当热点率为 0.4 和 0.5 时网络性能变得很差.相比两种网络拓扑,Fat Tree 网络抵抗热点通信的不良影响能力更强,这可以解释为 Fat Tree 中的多条冗余链路在一定程度上分担了热点消息的负载.



(a) Torus (b) Fat Tree

Fig.8 Latency curves under hot-spot traffic

图8 热点通信模式下的负载-延迟曲线

3.2.4 3D FFT 算法 2D 任务分解下的消息延迟

快速傅里叶变换(FFT)自 Cooley 和 Tukey 提出以来,已经成为众多科学领域中应用最为频繁的数值方法之一.针对三维数据集的三维快速傅里叶变换(3D FFT)已被使用在分子动力学模拟、湍流系统模拟等实际应用中,此类应用的计算规模往往很大,需要在大规模并行计算机上实现,这时网络的延迟将难以忽略.此外,由于 FFT 算法自身的特点,计算节点间的通信过程往往会占据程序运行总时间的一半以上,因此,研究 3D FFT 并行实现中消息的延迟具有实际意义.

假设进行 3D FFT 的三维数据集合是等长的,设 $A_{x,y,z}$ 为 3 维 $N\times N\times N$ 元复数值阵列,其 3D FFT 阵列 $A_{u,v,w}$ 定义为

$$A_{u,y,w}=\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\sum_{z=0}^{N-1}A_{x,y,z}\exp\left(-2\pi i\frac{ux}{N}\right)\exp\left(-2\pi i\frac{vy}{N}\right)\exp\left(-2\pi i\frac{wz}{N}\right),\quad 0\leqslant u,v,w\leqslant N-1 \tag{1}$$

目前 3D FFT 的并行实现方法主要是通过对数据集进行分解.相比数据集的 1D 分解,2D 分解具有更高的可扩展性,适合大型计算任务在大量节点的并行计算机上实现.2D 分解是将 3D FFT 的 $N\times N\times N$ 数据集分解到二维的 $P=P_{row}\times P_{col}$ 个进程矩阵中进行并行计算.每个进程包含 $N^3/P_{row}\times P_{col}$ 个数据,总的计算过程包括以下 5 步:

(1) 每个进程进行 $N^2/P_{row}\times P_{col}$ 个独立的x-维FFT;

(2) 每一行的 P_{col} 个进程之间进行两两通信;

(3) 每个进程进行 $N^2/P_{row}\times P_{col}$ 个独立的y-维FFT;

(4) 每一列的 P_{row} 个进程之间进行两两通信;

(5) 每个进程进行 $N^2/P_{row}\times P_{col}$ 个独立的z-维FFT.

第 1、3、5 步中每个计算节点利用自己的数据独立进行一维的 FFT 计算,不需要与其他节点进行通信;第 2、4 步中每个计算节点都进行一对多通信,与网络中特定的节点集合交换数据.因此,模拟 3D FFT 的消息延迟,只需考虑第 2、4 步的通信过程.

第 2 步中编号为 i 的进程 $(0\leqslant i\leqslant P_{row}\times P_{col}-1)$ 与编号为

$$\left\{\left\lfloor\frac{i}{P_{col}}\right\rfloor\cdot P_{col},\ \left\lfloor\frac{i}{P_{col}}\right\rfloor\cdot P_{col}+1,...,\ (i-1),\ (i+1),...,\ \left(\left\lfloor\frac{i}{P_{col}}\right\rfloor\cdot P_{col}+P_{col}-1\right)\right\} \tag{2}$$

的 $P_{col}-1$ 个进程进行通信,传递数据的量都相等.可以将第 2 步的通信过程近似为进程 i 产生的消息以等概率选取上述进程集合中的任何一个进行通信.

第 4 步中编号为 i 的进程 $(0\leqslant i\leqslant P_{row}\times P_{col}-1)$ 与编号为

$$\left\{ (i \bmod P_{col}),\ (i \bmod P_{col})+P_{col},...,\ (i-P_{col}),\ (i+P_{col}),...,\ \left[(P_{row}-1)\cdot P_{col}+(i \bmod P_{col})\right]\right\} \tag{3}$$

的 $P_{row}-1$ 个进程进行通信,传递数据的量都相等,同样可以将第4步的通信过程近似为进程 i 产生的消息以等概率选取上述进程集合中的任何一个进行通信.

根据具体的网络拓扑,需要将上述进程一对一映射到网络中的多个计算节点.对同一个并行应用选择不同的进程到计算节点的映射策略,通信开销可能存在较大差别,近年来已有研究者提出的“基于拓扑感知的任务分配”[12]正是着眼于这个问题.这里,3D FFT 进程映射仅考虑最简单的方法,即将每个进程映射到与其具有相同编号的计算节点上,前提条件是进程个数和网络中计算节点的个数相等.图 9 和图 10 分别是 2D 进程集合向 3D Torus 网络和 Fat Tree 网络节点映射的示意图.

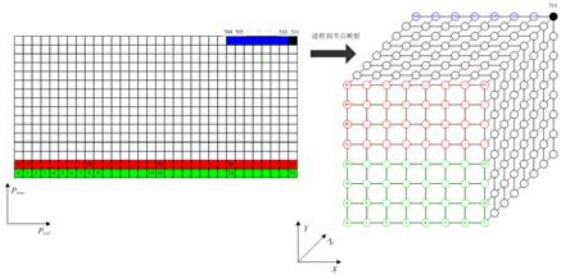


Fig.9 2D process set to nodes of 3D Torus mapping

图9 2D进程集合向3D Torus网络节点的映射

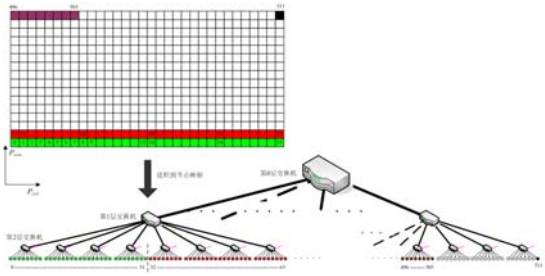


Fig.10 2D process set to nodes of Fat Tree mapping

图10 2D进程集合向Fat Tree网络节点的映射

研究中模拟了两种网络规模(512 节点,4 096 节点)下 Torus 和 Fat Tree 分别运行 3D FFT 的 2D 任务分解并行算法的消息延迟.假设网络中各个节点独立地以 Poisson 过程产生消息,Torus 网络使用维度优先路由算法,Fat Tree 网络使用模拟器默认的“Nearest Common Ancestor Random”路由算法.图 11、图 12 分别是消息长度为 32 微片时,两种网络 3D FFT 第 1 阶段和第 2 阶段的消息延迟.

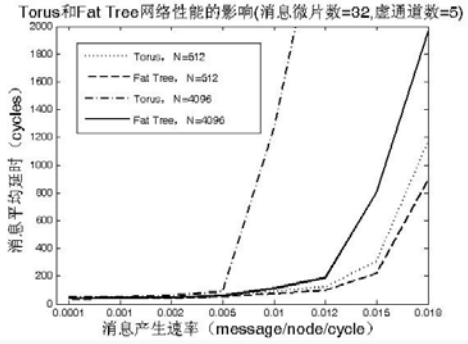


Fig.11 Message latency of the first stage in 3D FFT

图11 3D FFT第1阶段消息延迟

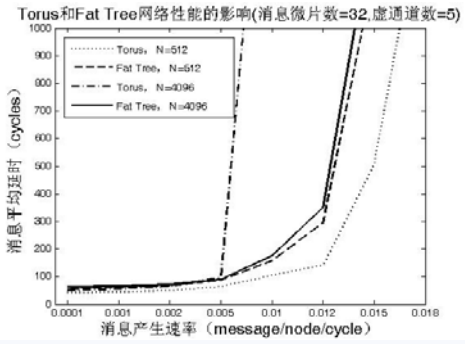


Fig.12 Message latency of the second stage in 3D FFT

图12 3D FFT第2阶段消息延迟

实验结果表明,对 512 节点情况 Torus 和 Fat Tree 网络第 1 阶段通信的消息延迟比较接近,当消息较短时(32 微片),Fat Tree 网络略优;4 096 节点情况下,Fat Tree 网络的消息延迟明显低于相同条件的 Torus 网络,并且与 512 节点时 Fat Tree 的延迟差距不大,这表明,当采用进程编号映射到网络中相同编号节点的策略时,Fat Tree 网络在第 1 阶段通信过程中具有优越性,并且消息延迟随网络规模的增长较为合理,具有良好的可扩展性.对 512 节点情况,Torus 网络第 2 阶段通信的消息延迟比较低;4 096 节点情况下,Fat Tree 网络的消息延迟明显低于相同条件的 Torus 网络,并且与 512 节点时 Fat Tree 的延迟差距不大.

通过上述分析,综合 3D FFT 并行算法的两阶段通信,当节点数目较少时(512),选用 Torus 较好;当节点数目较多时(4 096),应选择 Fat Tree 网络支持更高的可扩展性.

上述研究均考虑将每个进程映射到与其有相同编号的节点上运行,这往往不能保证相互通信节点物理距离上的临近性,特别是在 Fat Tree 网络中的第 2 阶段通信时表现尤为明显.节点间通信的空间局部性能够有效降低消息延迟,提升网络吞吐量,因此进一步的研究可以通过这种网络模拟的方法,考虑如何利用局部性,优化 3D FFT 等典型并行算法在各种网络下的映射策略,基于 Torus 网络,此问题学术界已有一定研究,而目前针对 Fat Tree 网络的研究仍然较少.

4 结 论

为了更加接近真实并行应用的网络负载,本文提出在之前研究中常常讨论的简单负载模型的基础上,增加考虑多种消息注入模式混合、局部通信、热点通信、长短消息混合分布等因素的复杂负载模型.扩展设计一个微片级网络模拟器,以支持模拟这种复杂负载模型下的互连网络性能.

模拟比较 Torus 与 Fat Tree 网络在不同负载下的性能,以及 3D FFT 算法 2D 任务分解下的消息延迟.实验结果表明:

(1) 目的地均匀负载模式下,对于大规模网络,Fat Tree 的性能往往优于 Torus;
(2) 并行应用的局部通信能够提高网络性能;对于 Torus 网络,局部通信率比局部通信区域半径对消息延迟的影响更加明显;Fat Tree 网络则相反,局部通信区域层数对消息延迟的影响更明显;
(3) 热点通信会大幅度增加消息平均延迟,与 Torus 网络相比,Fat Tree 网络抵抗热点通信不良影响的能力更强;
(4) 当采用进程编号映射到网络中相同编号节点的策略时,Fat Tree 网络在第 1 阶段通信过程中具有优越性;512 节点 Torus 网络第 2 阶段通信的消息延迟比较低;4 096 节点 Fat Tree 网络的消息延迟明显低于相同条件的 Torus 网络.

上述研究为具有局部通信、热点通信等复杂通信特点的并行应用提供了一种有效预测消息延迟和网络最大吞吐量的方法.模拟实验的数据及分析,可以有效支持大规模并行计算机互连网络的设计,提升并行程序执行效率.

References:

[1] Bhatele A, Kale LV. An evaluative study on the effect of contention on message latencies in large supercomputers. In: Proc. of the Workshop on Large-Scale Parallel Processing (IPDPS 2009). Rome, 2009.

[2] Duato J, Yalamanchili S, Ni L. Interconnection Networks: An Engineering Approach. Los Altos: Morgan Kaufmann Publishers, 2002.

[3] Johnson KL. The impact of communication locality on large-scale multiprocessor performance. In: Proc. of the 19th Annual Int'l Symp. on Computer Architecture. Queensland, 1992. 392–402.

[4] Kim JH, Chien AA. Network performance under bimodal traffic loads. Journal of Parallel and Distributed Computing, 1995,28(1): 43–64.

[5] Loucif S, Ould-Khaoua M. Performance analysis of deterministically-routed bi-directional torus with non-uniform traffic distribution. Future Generation Computer Systems, 2009,25(5):489–498.

[6] Hu K, Wang Z, Jiang S, Yin BL. A performance model of *k*-ary *n*-code under communication locality. Journal of Computer Research and Development, 2011,48(11):2083−2093 (in Chinese with English abstract).

[7] Dally WJ, Towles B. Principles and Practices of Interconnection Networks. San Francisco: Morgan Kaufmann Publishers, 2003.

[8] Culler DE. Parallel Computer Architecture: A Hardware/Software Approach. San Francisco: Morgan Kaufmann Publishers, 1997.

[9] Xu CF, Che YG, Wang ZH. Research on parallel performance simulation of large scale parallel computer. Computer Science, 2009,36(9):7−10 (in Chinese with English abstract).

[10] Bolch G, Greiner S, Meer H, Trivedi KS. Queueing Networks and Markov Chains. John Wiley and Sons, 1999.

[11] Pardo F, Boluda JA. SimuRed: A flit-level event-driven simulator for multicomputer network performance evaluation. Computers and Electrical Engineering, 2009,2

[12] Jagode H, Hein J. Custom assignment of MPI ranks for parallel multi-dimensional FFTs: Evaluation of BG/P versus BG/L. In: Proc. of the 2008 Int'l Symp. on Parallel and Distributed Processing with Applications. Sydney, 2008.

附中文参考文献:

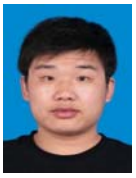
[6] 胡凯,王哲,蒋树,尹宝林.*k*-元 *n*-立方体网络局部通信模式下的性能模型.计算机研究与发展,2011,48(11):2083−2093.

[9] 徐传福,车永刚,王正华.大规模并行计算机系统并行性能模拟技术研究.计算机科学,2009,36(9):7−10.

胡凯(1963－),男,湖南长沙人,博士,副教授,CCF 会员,主要研究领域为高性能计算,嵌入式系统.

王哲(1984－),男,博士生,CCF 学生会员,主要研究领域为分布式,并行计算.

陈陆佳(1987－),女,硕士生,CCF 学生会员,主要研究领域为分布式,并行计算.



蒋树(1986－),男,硕士生,主要研究领域为并行计算,AADL.