ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, Supplement, December 2009, pp.23−33

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

一种海量数据流应用并行优化模型*

孙小涓[1+], 孙凝晖[2], 雷 斌[1]

[1](中国科学院 电子学研究所 空间信息处理与应用系统技术重点实验室,北京 100190)

[2](中国科学院 计算技术研究所 计算机系统结构重点实验室,北京 100190)

A Parallel Optimization Model for Massive Data Stream Application

SUN Xiao-Juan[1+], SUN Ning-Hui[2], LEI Bin[1]

[1](Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Key Laboratory of Computer System and Architecture, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

+ Corresponding author: xjsun@mail.ie.ac.cn, http://www.ie.ac.cn

Sun XJ, Sun NH, Lei B. A parallel optimization model for massive data stream application. *Journal of Software*, 2009,20(Suppl.):23−33. http://www.jos.org.cn/1000-9825/09004.htm

Abstract: While computing is entering a new phase in which CPU improvements are driven by the addition of multiple cores on a single chip, rather than higher frequencies. Parallel processing on these systems is in a primitive stage, and requires the explicit use and knowledge of underlying thread architecture. Based on the features of massive data stream application, this paper proposes three-level pipelining programming model of multithreading system, which realizes the new synchronization mechanism with no contention of shared structures and is capable to provide differential service for data streams. Then the paper applies the new model to remote sensing information processing system and backbone network intrusion detection system, and evaluates the improved system on several multicore platforms. In performance analysis, the optimized effects of backbone network intrusion detection system are evaluated in several aspects of throughput scalability on both SPARC T1 and x86 platforms, the impacts of different multithreading mapping methods on throughput, and the comparison of response time and service quality before and after optimization. The experimental results show that the system throughput has good scalability on both platforms, the values of response time are greatly improved and the prioritized streams achieve better response time with the differential service mechanism.

Key words: massive data stream; multicore; parallel optimization; multithreading programming model; network intrusion detection

摘 要: 计算进入了多核时代,处理器的发展不再由更快的主频带动,而是依靠增加片上的多个核心.但是,对于高性能应用来说,多核平台的并行处理由于缺少适合的并行程序开发工具还处于初始阶段,对应用的优化需要对底层线程结构的深入了解和正确使用.从海量数据流应用的特点出发,提出了三级流水多线程模型,它的线程同步机制没有竞争,并且实现了不同特征数据流的差别服务.然后,在遥感图像处理和骨干网网络入侵检测系

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2006AA01A102 (国家高技术研究发展计划(863))

Received 2008-07-01; Accepted 2009-04-02

统设计中,应用了海量数据流应用模型,并在多个多核平台下对骨干网网络入侵检测系统进行了性能评价.对 SPARC T1 平台上的线程映射方法进行研究,测试了不同映射方法的性能,并归纳出应用在体系结构方面的特征;采用 Sun SPARC T1 架构 8 核 32 线程服务器和曙光 x86 架构 8 处理器 16 核服务器对系统吞吐率进行了测试,实验结果都表现了良好的可扩展性;使用真实骨干网络流量记录文件回放产生的模拟流量,对比测试了模型应用前后数据流的服务时间,改进系统的响应时间获得了显著的提高;针对系统连接数大、负载重和处理多样性的特点,采用基于探针流的采样算法准确测试了在精确预测 IP 网段策略下系统的服务质量,同时也测试了增加服务质量优化后系统的延迟开销,实验结果表明,系统在引入较少延迟下提高了数据流的服务质量.

关键词: 海量数据流;多核;并行优化;多线程模型;网络入侵检测

随着多核计算机的发展,它提供给服务器大规模计算的能力.但是,硬件的大规模并发加大了软件开发的困难.多核处理器的发展给应用程序的开发带来了新挑战,它要求应用程序通过并行的多任务来执行.虽然一些典型应用基准测试[1],包括网络、数据库、密码等方面,已经证明在多核平台上具有很大的性能提高,但不是所有的已有应用都能很容易的获得在多核平台上的性能提高.

海量数据流应用具有数据量大、大规模并行和计算密集的特点,遥感图像处理系统和骨干网网络入侵检测系统都是海量数据流应用.遥感图像处理系统对用于地球资源研究的遥感数据进行采集和分析,获得影像、图形、统计表、GIS 数据库等处理结果.骨干网入侵检测系统是位于骨干网接口的入侵检测系统,它的基本功能是实现对网页、邮件、下载文件的实时检测,并对传输的病毒、色情等有害信息进行记录和报警,防止有害信息的传播和扩散.对遥感图像实时处理来说,数据流是由一次接收记录的格式化数据文件分解的所有景数据,而对于骨干网网络入侵检测来说,数据流是属于同一连接的所有报文.两者的数据单元分别是景数据和报文,不同在于景数据的处理可以不按序进行,而报文的协议重组处理必须按序进行.本文以此为研究背景,提出了一种适用于海量数据流应用的并行软件模型,并应用于实际系统中,获得了良好的性能.

从现有遥感图像处理系统和骨干网网络入侵检测系统的运行现状中,可以分析海量数据流应用面临的两大挑战:

一个是需要更高的处理能力,对计算机系统来说,虽然体系结构、大规模集成电路等领域的进展已经使得今天的计算资源非常丰富,但是在海量数据流应用中仍然显得十分有限.对于遥感图像处理,一方面,数据量较大,空间分辨率、时间分辨率、光谱分辨率和辐射分辨率不断提高,单幅遥感图像的数据量已达到数百兆,另一方面,遥感应用不断提出新的需求,很多领域,如气象预报、灾难监测等,需要对遥感图像进行快速及时甚至实时的处理,因此,对内存和计算资源的需求较大.对于骨干网网络入侵检测,根据 Gilder 定律(通信带宽每 6 个月翻一番,网络带宽的增长与 CPU 性能相比,至少要快 3 倍),带宽的增长导致了 TCP 连接数的大量增加,在 1Gbps 链路的广域网骨干路由器上一般会同时有几十万甚至上百万个 TCP 连接处于活跃状态,入侵检测系统的解码、格式分析以及扫描等处理也是计算密集的.

另一个挑战是需要提高有效性,现有的系统在海量数据流下,很难保证服务质量,无法区分不同数据流特征并对某些流量进行优先处理,用户需求得不到满足,用户所感知的性能也明显下降.在较大的系统负载下,遥感图像处理系统需要快速的响应时间,特别是用户感兴趣的关注地区和数据质量较高的地区,需要较快的处理速度.而在潮水般的网络流量下,骨干网网络入侵检测系统的待处理报文常常堆积,因此网络威胁事件的响应时间和反应速度变得不确定,甚至系统在过载时会错过做出反应的最佳时机.

本文将在以前研究[2]的基础上介绍新的研究成果.本文第 1 节介绍当前多核处理器发展和应用开发需求方面的相关工作.第 2 节抽象了基于数据流特征的三级流水多线程模型.第 3 节介绍把该模型应用在遥感图像处理和骨干网网络入侵检测系统的设计中.第 4 节对骨干网网络入侵检测系统进行了性能分析.最后,在第 5 节得出结论.

1 相关工作

随着单个芯片上晶体管数目的增加,处理器设计者需要提出下一代高性能处理器的体系结构.随着 CPU 与内存间速度差距的扩大,要提高处理器的速度除了需要设计有效的 Cache 系统之外,还需要通过并行化来隐藏访存延迟.在过去的 20 年里,处理器发展的主流是以 CPU 的主频提高驱动的,然而,功耗的问题已经开始限制处理器频率的增长速度.近年来高端处理器主要利用超标量(superscalar)技术来提高性能.超标量处理器在每个时钟周期发射多条指令到功能部件上执行,其目的是利用程序的指令级并行性(instruction-level parallelism,简称 ILP)来提高性能,但是单个程序的有限 ILP 导致了超标量处理器的资源利用率不高,对于下一代高性能处理器而言,开发并行性不应该仅限于单个程序内细粒度的 ILP,在许多实际工作负载中,存在多种形式的粗粒度的线程级并行性(thread-level parallelism,简称 TLP).研究人员提出多种利用 TLP 来提高处理器资源利用率的处理器体系结构,包括多线程处理器(multithreaded processor)、单片多处理器(chip multiprocessor,简称 CMP)和同时多线程(simultaneous multithreading,简称 SMT)结构.

越来越多的处理器生产商,已经发布了他们各自的 4 个或更多核的芯片,并提供相应的应用开发方法.Sun 已经发布了UltraSPARC T1 处理器,也曾被叫作“Niagara”[3,4],并将它使用在Sun Fire服务器[5]系统上.一个 Niagara 2 处理器,每核线程数目从 4 个增加到 8 个,从而可提供 64 个同时线程,可得到目前的UltraSPARC T1 处理器至少两倍的吞吐率,然而它们使用同样的功耗散热封装.本文的实验平台也选择了具有 8 核 32 线程 UltraSPARC T1 处理器的Sun Fire T1000 高性能服务器.Solaris作为多线程多进程操作系统,它提供了可扩展应用程序开发的多线程环境,如Solaris线程、POSIX(portable operation system interface)线程、多线程程序调试程序如TNF等.Cell高性能处理芯片的设计于 2005 年由IBM,Sony和Toshiba首次公开.Cell采用了与主流高性能处理芯片全然不同的片内分布式体系结构.它由一个相对比较简单的支持同时双线程并行的双发射 64 位 PowerPC内核(称为PPE)和 8 个SIMD(single instruction multiple data)型向量协处理器(称为SPE)构成.片内有一个高带宽的环状高速总线(EIB)把PPE,SPE及RAMBUS内存接口控制器(MIC)、FlexI/O外部总线接口控制器(BIC)连接起来.PPE主要负责控制并执行操作系统,SPE完成主要的计算任务.Cell处理芯片可在 4GHz频率下工作,其宣称的峰值浮点运算速度为 256GFLOPS.Cell的软件层次实现了两个级别的并行:SIMD和并行任务.实现并行性在计算方面体现为多核、多线程、多LS(local store)访问,在通信方面体现为DMA、总线带宽、流量控制、共享内存/消息传递、同步.Cell提供了全系统的、低开销的应用编程模型.选择适合的程序模型,可以在获得高性能的同时降低开发成本,程序结构抽象有助于提高生产力,也可以混合几种程序模型.

应用程序并行化是并行处理领域研究的热点问题,也是高性能计算需要解决的问题之一,这个问题由来已久,然而始终未能得到很好的解决.一般的并行化方法有两种:一种是程序员显式使用并行编程技术来开发应用程序,由于受限于现有的语言模型的局限性和求解问题的复杂度,这种方法没能得到大规模的推广;另一种就是依赖于并行编译器自动地将程序并行化.传统方法主要依靠并行编译器将程序编译成多个小的并行程序,比如斯坦福大学的SUIF编译器[6]等,已有研究成果主要应用在科学计算,不能适应于如桌面应用、多媒体及服务器等更广泛的领域.而一些软件事务内存(transactional memory)技术方面的研究努力简化并行应用开发,以获得在高度并行的系统结构下的性能扩展.这些利用并行编程环境的技术,和从头用一个并行执行模型来重写代码相比效果有限,一些针对应用特点重写代码的成功例子覆盖了从科学计算[7,8]到多媒体应用[9]的不同领域应用.

在协议处理应用领域的并行开发,一般也是利用多线程来实现处理的并行化.协议处理的并行策略根据应用的不同有很多类型,一般可以分为3种:基于层次的并行[10],即根据协议层次划分并行单位,实现起来清晰简单但是并行度受协议层数的制约,且由于协议层次实现复杂程度不同容易产生负载不均衡的情况;基于包的并行[11],即每个处理器或线程处理一个数据包,能够加速相同或不同连接的包处理速度,这种策略对于无状态协议处理有较好效果,但对于像TCP这样有状态协议来说相同连接内的数据包处理会受制于共享数据的互斥;基于连接的并行[12],以连接构成并行单位,因为实现简单,能够较好地挖掘连接并行性,且资源同步开销小,所以应用较广.除以上 3 种之外,还有基于模块或者函数的并行,即以函数模块为单位并行处理,但应用较少.

2 海量数据流应用模型

并行计算的核心问题是如何将数据进行划分,如何控制并发任务的执行,和并发任务间采用什么样的通信同步方式.为了解答这些问题,需要对数据流应用特点进行抽象.

我们提出的软件模型把整个工作流分成 3 个上下文,与T_{1n},T_{2n}和T_{3n}对应,用独立线程并行第 2 个和第 3 个上下文.如图 1 所示,数据流按照一定规则分发到多个待处理(P)队列上.在数据流单元准备好之前,如在TCP连接建立之前,它们可以被缓存在初始(S)队列.之后,数据单元按照一定策略加入一个工作(W)队列,模型使用很多W队列来满足不同服务级别的处理需求,而需要改变工作队列或将被放弃的数据单元被加入重分发(R)队列.各线程的任务划分如下:

(1) T_{1n}:获取完整的数据单元,将数据读取到内存中,通常包含磁盘I/O或者网络I/O操作.

(2) T_{2n}:初始化数据流工作区,创建数据流的数据结构,给基本字段赋值,进行数据有效性检验,完成必需的数据准备工作.

(3) T_{3n}:对数据单元进行处理.

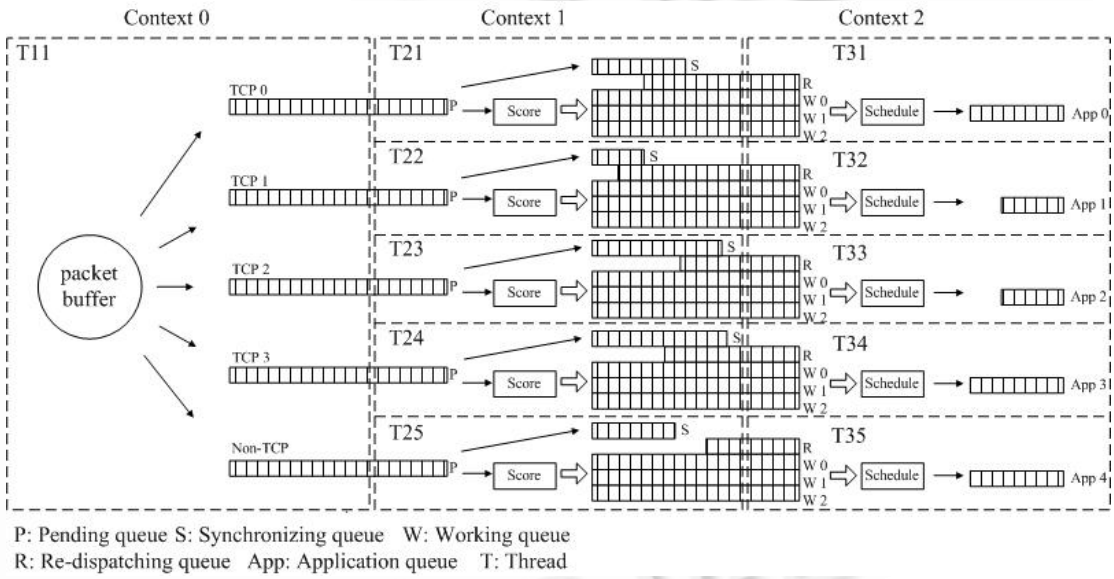


Fig.1 Multithreading model

图 1 多线程模型

该模型的特点是采用无锁同步通信方式的三级流水结构,可以实现不同数据单元的差别处理:第一,报文在不同的处理队列单向流动,并依照它们加入队列的时序被依次处理,采用无锁同步方式消除了线程间对共享队列的竞争,降低了消耗在互斥锁上的系统时间.第二,串行工作流可以分成几个可并行的上下文,每个上下文可以把队列作为输入和输出的管道.一个上下文具有独立的处理功能,可以分配单独的核完成.由上下文组成的流水结构适合运行在高度并行的多核系统.第三,通过很多 W 队列可以满足不同服务级别的处理需求.

3 模型应用

3.1 遥感图像实时处理

现有的遥感图像处理系统,以星载 SAR 信号处理系统为例,处理流程分为 3 个阶段.首先,系统源源不断地获取从传感器记录并传回地面系统的原始格式化数据文件,从原始格式化数据文件中读取数据,从辅助数据分离出成像处理必需的 GPS 数据和姿态数据,并将分散的辅助数据单元进行数据校验和误码纠错处理,整合成一个连续的辅助数据文件;其次,基于处理后的辅助数据对读取的原始数据进行数据校验和误码纠错工作,生成纠错

后的数据文件;然后,根据数据质量和被关注程度等因素,给分景后的数据文件指定处理优先级,生成这一景的标准格式图像.

现有系统在分布式计算环境上进行遥感信息处理,各处理节点使用机群文件系统进行数据共享,各节点通过消息通信机制控制并行任务的执行.流程控制节点将处理流程的前两个阶段调度到预处理计算节点,处理完成后,流程控制节点再调度成像处理计算节点完成第 3 个阶段,生成标准图像.我们将数据流应用模型应用于遥感信息处理中,系统独立运行在一台多核服务器上,各线程任务划分如下:

(1) T_{1n}线程负责格式化基本数据单元读取;

(2) T_{2n}线程负责分离数据单元的辅助数据,进行数据有效性处理,形成景数据;

(3) T_{3n}线程负责从景数据生成标准图像产品.

3.2 骨干网网络入侵检测

骨干网网络入侵检测系统的工作流程主要经过 3 个阶段,分别是报文接收、协议重组和内容过滤.在协议重组阶段,依次处理的是 IP 地址过滤、TCP 连接建立前的 3 次握手机制和建立后的数据排序以及应用协议的分析.报文可以自然地以 TCP 流为单位并行处理.

现有骨干网网络入侵检测系统的并行结构采用主从线程结构,报文由主线程分发到多个从线程处理.主线程通过锁同步机制与从线程通信,主从线程间同步开销很大.报文按接收时间顺序处理,不同连接的并发处理难以实现,更没有根据用户需求优先或延迟处理某些流的能力.我们将数据流应用模型应用于现有系统中,模型各线程任务划分如下:

(1) T_{1n}线程负责报文获取和IP分析;

(2) T_{2n}线程负责TCP的 3 次握手;

(3) T_{3n}线程负责TCP传输数据的排序、应用协议重组和内容过滤.

4 性能分析

本节将在吞吐率、响应时间和服务质量指标上对使用海量数据流应用并行优化模型的骨干网网络入侵检测系统进行性能评价.原有系统采用pthread多线程结构,报文由主线程分发到多个从线程处理,通过IP包头源和目的地址的哈希计算,分发规则确保同一连接的报文被指派到同一从线程.在以前的研究[2]中,分析了现有主从结构多线程骨干网网络入侵检测系统的性能瓶颈,主要表现在主线程和从线程共享数据互斥锁的系统同步开销大,报文在从线程处理队列堆积,等待服务时间较长.改进系统仍使用pthread实现多线程并行优化模型,运行在Linux和Solaris操作系统上.前文初步对比测试了改进系统在系统开销和吞吐率的提高,证明了改进策略的有效性.

本节将进一步通过实验分析系统性能:在不同架构多核平台上测试了系统吞吐率的可扩展性,研究了线程物理结构映射方法对吞吐率的影响;使用真实骨干网络流量记录文件回放产生的模拟流量,测试了模型应用后系统响应时间的提高;构建了 TCP 连接样本库,采用基于探针流的测试方法,对比测试了模型应用前后系统的服务质量.

4.1 吞吐率

4.1.1 不同多核平台的可扩展性

(1) SPARC T1 架构下的系统吞吐率

改进系统的吞吐率在SPARC T1 架构的 8 核系统下具有良好的可扩展性,线程映射方法是分别使用更多个核运行T_{2n}或T_{3n}线程,记作*xt*/*yc*,表示使用*y*个核,每个核用*x*个线程.同样的标识也出现在后续的测试中.测试结果如图 2 所示,报文吞吐率最高可达每秒 25.9 万,流吞吐率最高可达每秒 1.4 万.由统计结果来看,平均报文大小约 500 字节,因此系统可处理超过 1Gbps的骨干网流量.

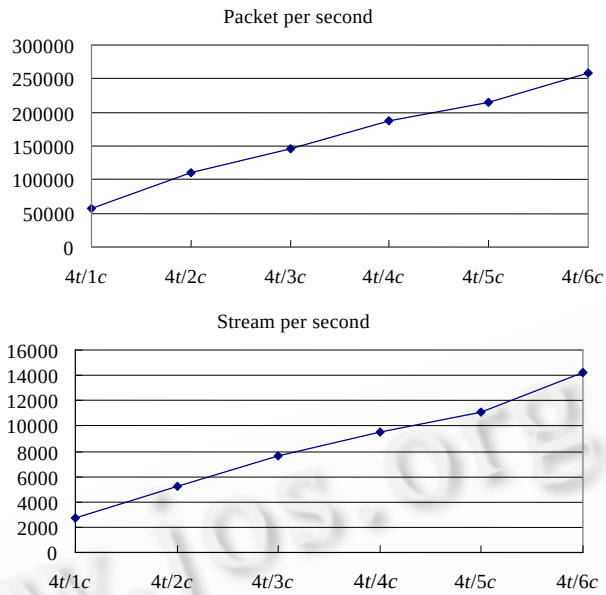


Fig.2 Scalability of throughput on SPARC T1 architecture

图 2 SPARC T1 架构下系统吞吐率的可扩展性

(2) x86 架构下的系统吞吐率

改进系统在 SPARC T1 架构下具有较好的性能测试结果,在其他多核平台的扩展性表现怎样?目前,双核 CPU 在高性能服务器广泛采用,我们选择了 x86 架构下 8 路双核处理器的 SMP 服务器平台,对改进系统吞吐率进行了测试.实验平台使用曙光 A950r-F 高性能服务器,它具有 8 个 AMD Opteron 8218 双核处理器,主频 2.6GHz,内存 16GB,系统总共有 16 个核.

将同一队列的二、三级处理线程分别绑定在同一处理器的两个核心(如图 3 所示,其中P0~P7 为处理器编号),性能测试如图 4 所示.可以看出,报文吞吐率随处理器数目增加接近线速增长.其中由于 6p的T_1线程分布在两个CPU,trace文件读带宽较高,所以性能较 5p略好.由测试结果可知,在 16 核平台上,报文吞吐率最高可达每秒 85 万,以平均报文大小约 500 字节计算,系统可处理超过 3Gbps的骨干网流量.

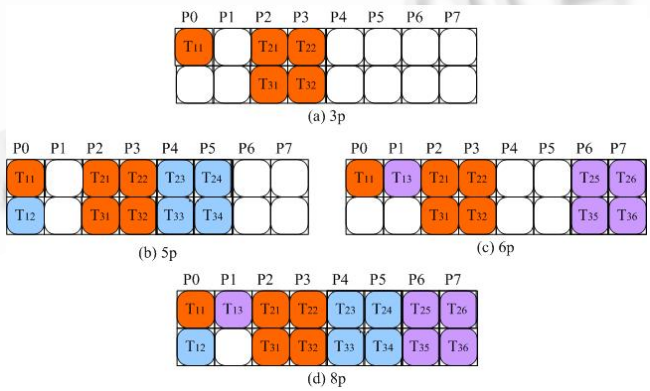


Fig.3 Multithreading mapping methods of x86 sixteen-core platform

图 3 x86 架构下 16 核系统多线程映射方法

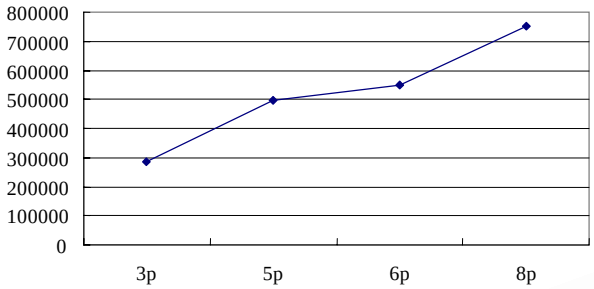


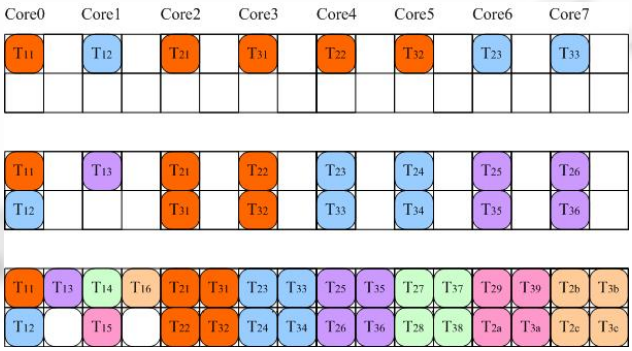
Fig.4 Throughput of x86 sixteen-core platform, packets per second (Units:pps)

图 4 x86 架构下 16 核系统报文吞吐率(单位:pps)

4.1.2 物理结构映射对系统吞吐率的影响

物理结构映射方法可能会影响数据处理的性能, 对于线程级并行体系结构,面临的问题有如何将线程与众多计算核心进行物理映射来达到最大的系统性能.一方面,SPARC T1 处理器核内的 4 个同步线程具有共享的流水线,处理器可以无切换开销地调度核内的其他线程使用流水线,来隐藏当前线程的访存和流水停顿带来的延迟.但是,它们共享指令和数据一级高速缓存(L1),并发的多个线程需要避免 Cache 冲突不命中以获得较好的性能.另一方面,共生是指多个作业在多线程架构上同时执行.作业在多核多线程处理器上会在共享的系统资源上互相发生冲突.吞吐率实际上会根据运行作业集的作业共生(或相处)的好坏增长或下降.因此,哪些作业一起调度可能会大不相同.考虑共生的调度策略可能带来更高的吞吐率和响应时间.

在实验中,根据多线程应用模型的特点,逐个测试了所有可能的调度实现,试图找到最佳的方案.对于核内同时线程映射(如图 5 所示),可以把核内使用的线程数从 1 个增加到 2 个、4 个,来测试它们的报文吞吐率和流吞吐率.实验结果显示,当线程数不断增加时,两个吞吐率接近线性增长,因此可以尽量利用核内同步线程资源,核内线程之间的不利影响并不明显.对于线程间共生映射(如图 6 所示),T_{1n},T_{2n}和T_{3n}线程执行不同的并行任务,可以通过绑定同类线程和不同类线程在同一个核内,分别测试同类线程和不同类线程的共生映射方法的系统吞吐率.测试结果显示,不同的绑定方案带来几乎相同的系统性能,平均单核报文吞吐率均在 52 000 左右,流吞吐率均在 2 500 左右,说明不同功能线程共生良好,不存在明显的资源冲突.



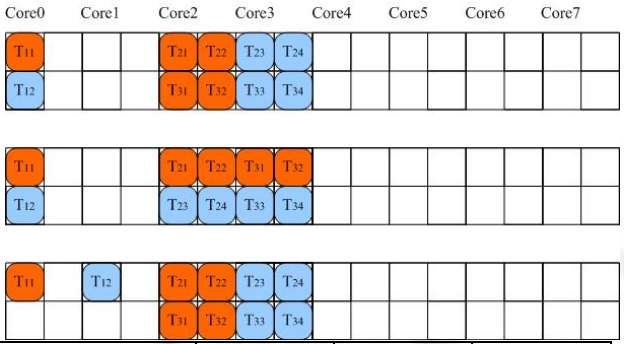
Core used	1*t*/6*c*	2*t*/6*c*	4*t*/6*c*
Packet per second	83 000	175 000	261 000
Stream per second	5 100	8 900	14 400

Fig.5 SMT scalability of the multithreading model

图 5 多线程模型 SMT 的可扩展性

分析这两个实验结果的原因,系统在不同映射方法下所表现的特征是应用的数据处理特点引起的.虽然内存访问行为差别很大,但是SMT通过并发处理更多访存负载不命中,隐藏了访存延迟.在不同类型应用中,网络应用SMT隐藏延迟的能力是最强的[13,14].这与我们实验数据所体现的是一致的.虽然各线程划分任务不同,但在

处理流程中仍发生大量的访存不命中,使得核内共生线程频繁进入数据等待,让出指令运算部件,而网络数据通常只访问一次,数据访问空间局部性差,所以核内线程很少竞争有限的计算和Cache资源,不管是同类线程或不同线程在同一核内调度,系统吞吐率都无较大变化.因此,骨干网网络入侵检测系统的线程映射要尽量利用核内同时线程,可以达到成倍提高系统性能的目的.



Core used	4*t*/2*c*	4*t*/2*c*	4*t*/2*c*
Packet per second	104 000	101 800	106 600
Stream per second	5 090	5 100	4 890

Fig.6 Threading mapping methods of symbiosis

图 6 多线程模型线程间的共生映射方法

4.2 响应时间

服务时间测量了从连接起始报文被接收到连接被删除的一段时间.它体现了连接在系统中的逗留时间,也反映了系统的实时快速响应能力.

实验平台选用双 CPU 四核服务器,它有两个 AMD Opteron 1.8GHz 双核处理器.实验负载来自 2005 年从北京某 ISP 采集的骨干网流量片段,流量在 190Mbps 左右,约 100 万报文,通过记录文件回放和设置关键字列表,模拟网络威胁安全事件检测过程.

实验记录了发生安全事件的那些连接的服务时间,对照图 7(a)和图 7(b)可以发现,模型应用后系统的服务时间比原系统要快,几乎所有连接的服务时间都有所提高,统计显示,PDF 在 0.5s 附近取值最高,也就是较多的连接提高 0.5s 左右,而 CDF 取值 0.5 时大于 0.5s,也就是 50%的连接提高幅度超过 0.5s.在高速的网络环境中,在极短的时间内就可以处理大量的网络报文,减少报文堆积,这样幅度的性能改进极大地提高了系统服务质量,处理有效性和用户感知性能都得到了提高.

4.3 服务质量

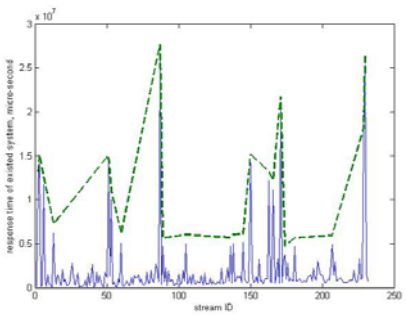
对于大规模的网络流处理,提前预知每个流的特征,跟踪它们的处理优先级定义,并测量每个流的服务时间比较困难,因为耗时的分析使得被测程序无法正常运行,所测数据也产生有很大偏差.所以,我们基于连接样本库基准测试集,采用探针流采样方法对服务质量优化系统进行性能评价,不但消除了测量误差的问题,而且对比测试也容易实现,优化效果可以评价.构建连接样本库是以 1:100 的比例抽样 40GB 连续记录文件、总计约 170 万个 TCP 连接,抽样的连接集合保持了原始数据的分布特征,通过统计分析具有极高的相似度.

实验测量的服务时间开始于连接初始报文到达时刻,终止于报文所属样本流删除时刻.测试系统记录了流到达时刻,流终止时刻和系统在此时的负载,如等待处理的报文数.实验通过对比流在两个不同优先级队列的服务时间,得到系统的服务质量优化效果.根据流量特征分析统计结果,一半的 TCP 流持续不足 3.03s,实验所采用的探针流是连接样本库中 3s 左右大小的流.设定打分策略是优先处理某些网段流量,调整优先处理的 IP 地址段,使得负载报文的优先和非优先比重相同,即不同优先级队列的到来报文速率相等.工作队列具有一个优先队列和一个非优先队列,调度策略简化为总是处理高优先级流.探针流随机的插入到达报文流里,一个探针流报文被插入的同时,会产生一个复制报文作为参照.如果该报文根据打分策略被分发到优先队列,那么其复制报文一

定被分发到非优先队列,反之亦然.

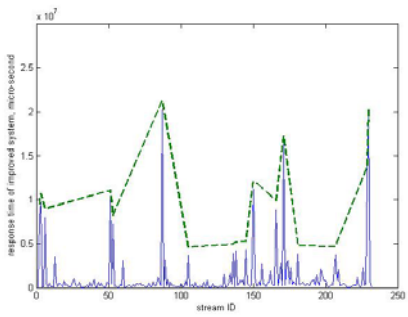
实验结果如图 8 所示,在不同系统负载情况下,绝大多数非优先流比优先流具有较长的服务时间,指定网段流量的服务质量得以保证.而且不管系统待处理负载有多少,优先流的服务时间稳定在 3s 左右,而作对比的非优先流服务时间波动较大,系统保证了优先流稳定的服务质量.

另外,无锁的同步机制和三级流水的多线程模型提高了系统吞吐率和响应时间,但是系统从两级主从结构[2]改变为三级流水结构增加了报文移入和移出工作队列的操作,也增加了系统的延迟.为了考察系统增加第 3 级的延迟,我们还对比了二级和三级结构中HTTP分析时间.对比实验结果显示,三级结构确实使绝大部分流的反应时间延长了,但是 80%的流具有很短的小于 161ms的延迟,与改进前系统平均 1 500ms的反应时间相比,这个延迟是可以接受的.



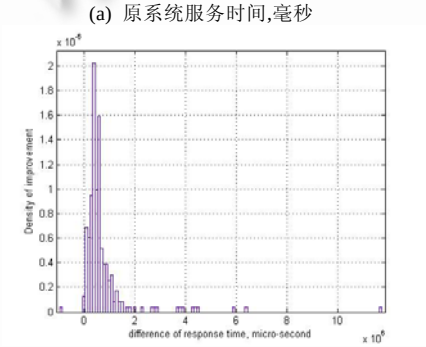
(a) Service time of existed system, ms

(a) 原系统服务时间,毫秒



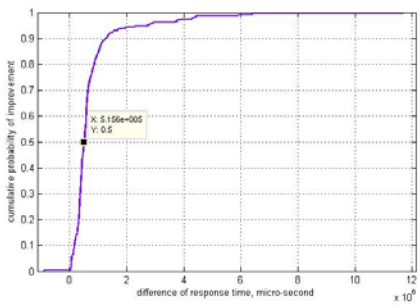
(b) Service time of improved system, ms

(b) 改进系统服务时间,毫秒



(c) PDF

(c) PDF 分布



(d) CDF

(d) CDF 分布

Fig.7 Comparison of service time before and after optimization

图 7 比较改进前后系统的服务时间

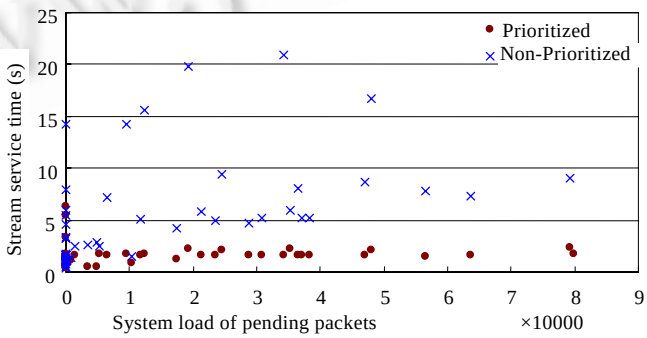


Fig.8 Comparison of service time in different priority queues

图 8 不同优先级队列服务时间比较

5 结 论

本文通过分析海量数据流应用特点,提出了适合多核平台的三级流水多线程模型,模型采用无锁队列的同步机制,可以实现不同特征数据流的差别服务.在对骨干网入侵检测系统的性能评价中,对 SPARC T1 平台上的线程映射方法进行研究,测试了不同映射方法的性能,得到应用在体系结构方面的特征;采用 Sun SPARC T1 架构 8 核 32 线程服务器的吞吐率,并在曙光 x86 架构 8 处理器 16 核服务器上测试了优化系统的可扩展性;使用真实骨干网络流量记录文件回放产生的模拟流量,对比测试了模型应用前后数据流的服务时间;针对系统连接数大,负载重和处理多样性的特点采用了基于探针流的采样算法,并准确测试了在精确预测 IP 网段策略下系统的服务质量,以及增加服务质量优化后系统的延迟开销.实验结果表明系统吞吐率表现良好的可扩展性,响应时间和服务质量都得到了显著提高.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是中国科学院计算技术研究所孙凝晖研究员领导的讨论班上的同学和老师表示感谢.

References:

[1] Sun Microsystem, Inc. Sun Fire T1000 and T2000 Servers Benchmarks. http://www.sun.com/servers/coolthreads/t1000/benchmarks.jsp

[2] Sun XJ, Sun NH, Chen MY. Optimization of B-NIDS for multicore. Journal of Computer Research and Development, 2007,44(10):1733−1740 (in Chinese with English abstract).

[3] Kongetira P, Aingaran K, Olukotun K. Niagara: A 32-way multithreaded SPARC processor. IEEE Micro, 2005,25(2):22−29.

[4] Sun Microsystem, Inc. OpenSPARC T1 Microarchitecture Specification. 2006. http://opensparc-t1.sunsource.net/specs/UA2005-current-draft-P-EXT.pdf

[5] Sun Microsystem, Inc. Sun Fire T2000 Server. 2007. http://www.sun.com/servers/coolthreads/t2000/

[6] Craig Z. Master/Slave speculative parallelization and approximate code [Ph.D. Thesis]. Madison: University of Wisconsin-Madison, 2002.

[7] Alarm SR, Barrett RF, Kuehn JA, Roth PC, Vetter JS. Characterization of scientific workloads on systems with multi-core processors. In: Proc. of the IEEE Int'l Symp. Workload Characterization. San Jose, 2006. 225−236.

[8] Hongzhang S, Erich S, Ji Q, David HB, Kathy Y. Performance modeling and optimization of a high energy colliding beam simulation code. In: Proc. of the ACM/IEEE SC 2006 Conf. Tampa, 2006.

[9] Berekovic M, Stolberg HJ, Pirsch P. Multicore system-on-chip architecture for mpeg-4 streaming video. IEEE Trans. on CSVT, 2002,12(8):688−699.

[10] David DC. The structuring of systems using upcalls. In: Proc. of the 10th ACM Symp. on Operating Systems Principles. Washington, 1985. 171−180.

[11] Mats B., Per G. Locking effects in multiprocessor implementations of protocols. In: Proc. of the ACM SIGCOMM Symp. on Communications Architectures and Protocols. San Francisco, 1993. 74−83.

[12] David JY, Erich MN, James FK, Don T. Networking support for large scale multiprocessor servers. In: Proc. of the SIGMETRICS. 1996. 116−125.

[13] Jack LL, Susan JE, Joel SE, Henry ML, Rebecca LS, Dean MT. Converting thread-level parallelism into instruction-level parallelism via simultaneous multithreading. ACM Trans. on Computer Systems, 1997,15(2).

[14] Susan JE, Joel SE, Henry ML, Jack LL, Rebecca LS, Dean MT. Simultaneous multithreading: A foundation for next generation processors. IEEE Micro, 1997,17(5).

附中文参考文献:

[2] 孙小涓,孙凝晖,陈明宇.多核平台上 B-NIDS 的优化.计算机研究与发展,2007,44(10):1733−1740.

孙小涓(1980－),女,山东烟台人,博士,助理研究员,主要研究领域为高性能计算机体系结构,入侵检测系统,遥感信息处理系统.

孙凝晖(1968－),男,博士,研究员,博士生导师,主要研究领域为并行计算机体系结构,分布式操作系统,计算机系统性能评价.

雷斌(1977－),男,博士生,副研究员,主要研究领域为合成孔径雷达图像处理,遥感信息处理系统.