文章编号:1671-4598(2018)06-0176-04 DOI:10.16526/j.cnki.11—4762/tp.2018.06.045 中图分类号:TP309 文献标识码:A

大数据环境下并行数据传输完整度控制方法

曾光辉,唐国强
(广州工程技术职业学院 信息工程系,广州 510900)

摘要:针对当前并行数据传输过程中节点拥塞严重、数据传输速率低、数据完整度低的问题,提出大数据环境下并行数据传输完整度控制方法;通过并行数据传输完整度控制原理,对传输中并行数据获取和传输拥塞度测量两大干扰因素加以分析;引用单服务窗混合制排队模型和栅格分析法分别对单节点和并行数据传输速率进行预控制,利用数据流优先级完成并行数据传输完整度控制方法的实现;实验结果表明,所提方法传输控制效果好,有效缓解了节点拥塞现象,提高了数据完整度。

关键词:大数据环境;数据传输;完整度;控制

Parallel Data Transfer Integrity Control Method in Large Data Environment

Zeng Guanghui, Tang Guoqiang
(Department of Information Engineering, Guangzhou Institute of Technology, Guangzhou 510900, China)

Abstract: Aiming at the problems of serious congestion, low data transmission rate and low data integrity in the process of parallel data transmission, a method of integrity control for parallel data transmission in large data environment is proposed. Based on the principle of integrity control of parallel data transmission, two interference factors of parallel data acquisition and transmission congestion measurement in transmission are analyzed. Reference single service window mixed queuing model and method were pre control of single node and parallel data transmission rate, using data flow priority to realize parallel data transmission method of complete control. The experimental results show that the proposed method has good transmission control effect, effectively alleviates the congestion of nodes and improves the data integrity.

Keywords: large data environment; data transmission; integrity; control

0 引言

数据传输控制是数据有效获取、减少网络能耗、最终完成数据完整传输的重要保障[1]。无线传感网络具有多变性特点,在数据传输时易产生无线链路数据传输误码率高,并行数据传输碰撞丢包率高,网络节点暴露受损导致数据交互异常率高等问题[2],另一方面,网络节点会因能量的损耗导致其失效,进而引发数据传输路径的无效[3]。节点数据存储能力有限,易出现数据流量异常增加,或数据包冲出缓存区丢失。网络节点的非对称性,也会导致信道的带宽分配不均衡[4]。综上所述,对数据传输完整度控制方法的研究成为了亟待解决的问题。在大数据背景下,海量数据多以并行方式进行传输,这也正是数据传输完整度控制方法的难点所在。实现数据的有效传输控制,也是提高网络带宽的利用率以及实现网络公平性中的必然需求。因此提出大数据环境下并行数据传输完整度控制方法,提高网络传输的稳定性及传输数据的完整度。

1 并行数据传输完整度控制原理

在无线网络并行数据传输的过程中,使数据完整度较低的,也就是丢包率现象严重的原因主要就是:传输节点拥塞致使数据包产生大规模碰撞[5-6]。所以解决上述问题就能够实现大数据环境下并行数据传输完整度的控制。具体原理如图1所示。

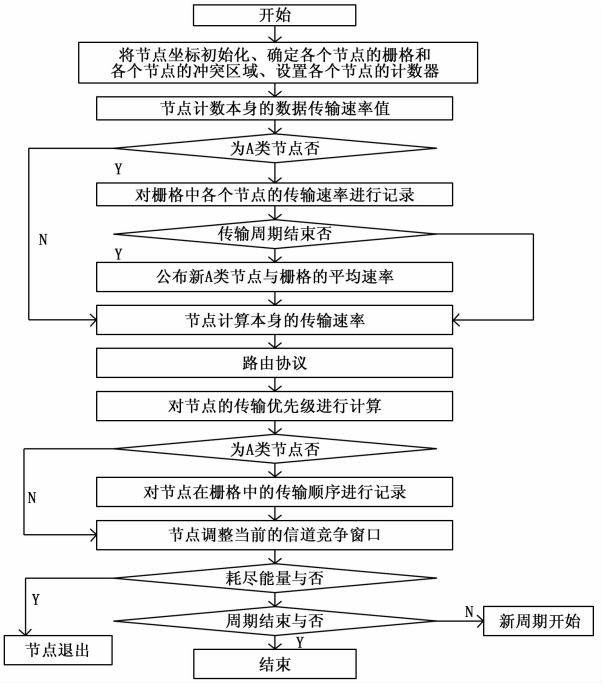

图1 并行数据传输完整度控制原理图

2 并行数据传输干扰分析

在研究并行数据传输完整性控制问题之前,需先对并行数据传输的干扰因素进行分析。在数据并行传输过程中,网络拥塞现象造成的数据传输速度低,数据传输控制较为困难等问题,都会对数据完整度控制造成严重影响。该部分对并行数据获取干扰和数据传输拥塞度测量干扰加以分析,以便更好地实

收稿日期:2017-10-13; **修回日期**:2017-10-29。

作者简介:曾光辉(1972-),男,湖南长沙人,硕士,副教授,主要从事智能信息处理方向的研究。

现并行数据传输完整度控制。具体分析如下：

2.1　并行数据获取干扰

为了减少在一个节点上的通信流量，各进程各自并行读取所需要的数据，从而加快读取数据的速率[7]。但这种获取数据的形式对集群硬件要求较高，集中的磁盘阵列设备可解决硬件要求高的问题，以便更有效的获取并行数据。

对并行数据进行获取，各进程间同时且并发的接收计算任务，在各自接收计算任务后，获取各自的数据集。详细过程如图 2 所示。

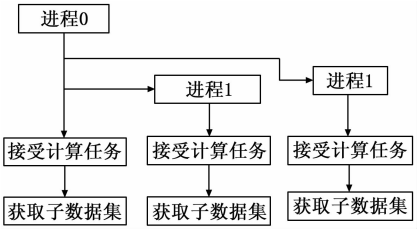

图 2　并行数据获取流程图

2.2　数据传输拥塞度测量干扰

对数据传输拥塞度进行测量用以分析每个传输节点的拥塞情况。首先对流入的数据包进行处理，按数据包速率平均分配给自身节点和其他子节点。通过这种节点局部公平速率分配的策略，可确保并行数据传输完整度控制的准确性。数据包的传输服务时间是从数据包到达 MAC 层算起的[8]，记录数据包的传输服务时间，传输间隔时间及传输到达时间，引用移动加权平均法计算数据传输平均服务时间和平均到达时间。数据传输拥塞度依据数据包平均传输服务时间和数据包平均到达时间求得。当平均到达时间比平均服务时间小时，数据传输拥塞度高，说明节点拥塞较为严重。反之，数据传输拥塞度低，节点拥塞程度小，数据传输速度快。对数据传输拥塞度的测量为并行数据传输完整性的控制做好充足的准备。

3　控制方法的实现

由于并行数据传输过程中的拥堵现象，经常出现于局部网络及其相关的多个节点上[9]。由此利用 CL－APTC 协议分别对单节点速率、系统级并行数据传输速率进行预控制，依据其结果对网络各节点实际传输速率加以控制，从而并行数据传输完整度控制方法，同时也缓解了数据传输过程中网络拥堵问题。

3.1　单节点速率的预控制

当单个节点的存储空间 $L < L_{\max}$ 时，CL－APTC 认为该数据节点不会出现拥堵的现象。当 $L_{\max} < L < m$，则说明数据节点在传输过程中有可能出现拥塞，应调节并行数据传输的速率，其中，L 代表数据节点目前所占用的空间，$L_{\max}$ 代表数据传输速率的调整阈值，m 代表节点最大的存储空间。如果节点的存储空间已经占满，也就是 $L = m$，那么节点传输的速率能够调整为 0。当新的周期 t 开始，通过周期 $t-1$ 栅格平均的传输速率预测周期 t 数据量的期望值 $E(t)$。

当 $E(t) < n^* L_{\max}$ 时，CL－APTC 认为数据传输不会出现系统级的拥塞现象，其中 n 代表某个栅格中的几点数量。当 $n^* m > E(t) > n^* L_{\max}$ 时，CL－APTC 认为网络中可能产生了系统级的拥塞，需要调整网络的栅格平均传输速率。依据用户的需求和环境情况，设置节点级和系统的速率权重，获得各个节点的最优输入和输出速率值：$\lambda^k_{real,i}(t)$ 和 $\mu^k_{real,i}(t)$。那么在实际的应用中，如果下游的节点总输出速率与上游的节点输入速率相等，节点 i 输出的数据由 $i+1$ 的数据输出决定。网络中的各个节点依据本身的输入与输出关系，获得其数据的转发速率，其中图 3 为速率的发送模型，详细介绍了网络数据传输速率的预控制和调整方式。

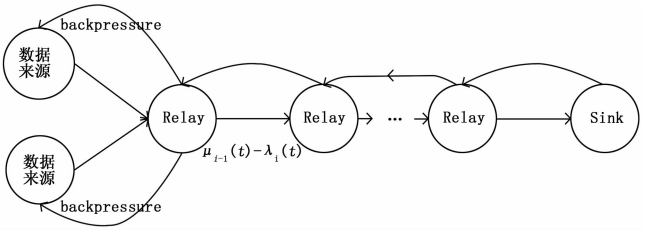

图 3　网络节点传输关系

根据图 3 的描述，得到对单节点速率的预控制和调整方法。因为网数据节点的存储空间是有限的[10]，通过 M/M/1/m（单服务窗混合制的排队模型）分析各个节点的传输速率，它的稳定状态如图 4 所示。

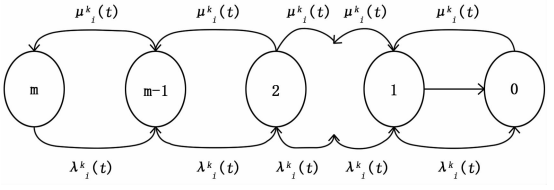

图 4　平稳状态图

根据图 4 可知，M/M/1/m 稳态方程为：

$$p_L = \frac{1-\rho}{1-\rho^{m+1}} \times \rho^L \tag{1}$$

式中，当 $\rho = 0.6$ 时，数据包的数量会比较缓慢地增加，系统会逐渐向最佳的状态发展。当 $\rho > 0.6$，系统会迅速至饱和状态，会导致数据包的溢出现象变得严重。当系统在初始状态时，依据 BS 单位时间数据量要求和 ρ 值的限制，获得数据源的采样速率。在 WSNs 中，如果节点缓存的占用量比其缓存空间的阈值 $L_{\max}$ 大时，数据节点也许会出现拥塞现象，则网络节点产生拥塞的概率为：

$$P_{con} = \sum_{L=L_{\max}}^{m} p_L = \sum_{L=L_{\max}}^{m} \left(\frac{1-\rho}{1-\rho^{m+1}} \times p_L \right) \tag{2}$$

根据上式能够获得 ρ、数据存储空间 L、网络节点拥塞概率间联系。由此依据 BS 要求的数据量，能够获得中间并行数据的转发速率。上述中提到当 $L_{\max} < L < m$，则说明数据节点有可能出现拥塞，应调节并行数据传输的速率。那数据节点本身的拥塞概率可表示为：

$$p^{con}_{\max} = \frac{(m-L_{\max})-(m-L)}{m-L_{\max}} = 1 - \frac{m-L}{m-L_{\max}} \tag{3}$$

历经一段时间，如果 L 重新回至最佳的区间内，也就是 $L \in [0, L_{\max}]$，则将保持 ρ 不变。如果 $L = m$ 且一直持续，节点 i 至 $i+1$ 的信道会出现严重的拥塞，i 利用 ACK 消息通告 $i-1$ 数据节点，然后迭代到数据的源节点。当 L 进入至最佳的区间，根据 ACK 方式来通告 $i-1$ 节点，并迭代到数据源。

当网络节点可能会产生拥塞时，通过 $\rho = (1 - p^{con}_{\max})\rho = ((m-L)/(m-L_{\max}))\rho$ 和 ACK 消息能够有效解除拥塞。综上

解决了节点级数据传输拥塞问题。不过网络拥塞的产生有空间关联性，由此还要考虑系统级的拥塞处理方法进行研究。

3.2 系统级并行数据传输速率的预控制

以并行数据传输平均速率调整的方便性为目的，利用前一个周期 $t-1$ 占用的局部存储资源 $E(t-1)$ 和平均的数据传输速率，对本周期 t 存储资源的占用量 $E(t)$ 进行预测，通过 $E(t)$ 和 $n*L_{\max}$ 的联系及 $E(t)$ 与 $E(t-1)$ 之间的比例，决定了周期 t 内的平均传输速率调整方式，进而避免数据传输过程中的局部拥塞现象。则系统级并行数据传输控制的具体流程如图 5 所示。

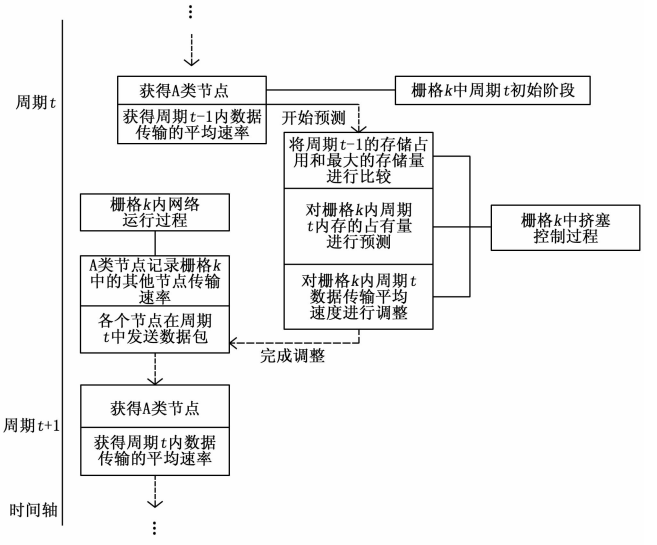

图 5 系统级并行数据传输控制流程图

由图 5 可知，当周期 t 初始化时，如果栅格 k 内活动的节点数量：$X(t)=n<N/k$，那么假设数据达到栅格 k 的概率和 Δt 为正比，可记为 $b_n*\Delta t$，到达两个或者两个以上数据概率为 $O(\Delta t)$。假设数据从栅格 k 中输出的概率和 Δt 为正比，记为 $d_n*\Delta t$，输出数据两个或者两个以上的概率为 $O(\Delta t)$。则为了分析栅格中数据的变化量，并获得 $P_n(t)$，也就是事件的概率可分解成：

1) $X(t)=n-1$，Δt 中输入的栅格数据量为 1，且概率为 $P_{n-1}(t)*b_{n-1}*\Delta t$；

2) $X(t)=n+1$，Δt 中输出的栅格数据量为 1，且概率为 $P_{n+1}(t)*d_{n+1}*\Delta t$；

3) $X(t)=n$，Δt 中没有数据量传输，也就是数据量并没有变化，则概率为 $P_n(t)\times[1-d_n*\Delta t-b_n*\Delta t]$。

综上所述，可以得到栅格中数据变化的期望值以及栅格内数据存储量变化的比例表达式分别为：$E(t)=E(t-1)\times e^t$ 和 $\delta=E(t)/E(t-1)=e^t$。当 $0\leqslant E(t)\leqslant nL_{\max}$，数据传输的平均速率差值保持周期 $t-1$ 不变，假设 $n*m>E(t)>nL_{\max}$，数据传输的平均速率差值应该减少，因此数据传输的平均速率差值应该调整为：

$$\lambda_{avg}^k(t)-\mu_{avg}^k(t)=(\lambda_{avg}^k(t-1)-\mu_{avg}^k(t-1))\times\frac{1}{\delta} \quad (4)$$

根据式 (4) 可得知 $E(t)$ 的求解公式为：

$$D(t)=E(t-1)*\frac{\lambda_{avg}^k(t)+\mu_{avg}^k(t)}{\lambda_{avg}^k(t)-\mu_{avg}^k(t)} \quad (5)$$

据上式可知，当数据输出平均速率和数据输入平均速率在频繁变化时，系统局部网络吞吐量会出现十分强烈的抖动，同时根据上述过程也解决了系统级数据传输拥塞问题，有效调整了数据传输的速率，提高了数据传输的完整度。

3.3 网络各节点实际传输速率控制

根据 3.1 和 3.2 的预控制调整结果，出于对节点级以及系统级数据传输速率状况的考虑，节点实际的传输速率可表示为：

t 周期数据输入的速率为：

$$\lambda_{real,i}^k(t)=\alpha\lambda_{avg}^k(t)+(1-\alpha)\lambda_i^k(t) \quad (6)$$

t 周期数据输出的速率为：

$$\mu_{real,i}^k(t)=\alpha\mu_{avg}^k(t)+(1-\alpha)\mu_i^k(t) \quad (7)$$

由上可知，历经节点级与系统级结合的数据传输速率预控制调节之后，网络各节点数据传输的速率根据式 (6) 和 (7) 来决定。这样能够综合地考虑整体和个体间的联系。当网络带宽和信道的质量资源较高时，可增加权重 α，使网络的整体资源得以高效使用；当某个节点数据的转发量较低，且子节点比较少时，可减少权重 α，使该节点本身拥有的资源得以充分利用，进而使整个网络运行的效率最高，从而控制并行数据传输完整度。

为保障数据传输的可靠性且减少节点数据传输的冲突，应减少数据包丢失率和优先级的需求，CL－APTC 协议对数据传输方案进行改善。利用数据流的优先级和等待时间大小，对目前的竞争窗口 CW 进行动态调整；各节点随着自身竞争信道的次数不断增加，相应地增加竞争至当前时隙的总体概率，也就是逐渐减少 CW 大小；当网络节点于周期 t 内，竞争至时隙且发送数据之后，它的竞争概率会降低到最小，一直到该周期结束。详细过程为：

假设，数据流 q 在节点 i 等待发送的时间为 $W_i^q(t_1,t_2)$，长度是 L_{data}^q，原始的优先级是 $p_{initial}^q(t)$，则数据发送的优先级是 $p_i^q(t)=p_{initial}^q(t)\times W_i^q(t_1,t_2)$，节点将 $\max(p_i^q(t))$ 当作预备发送的数据流。周期 t 初始化，节点将 $\max(p_i^q(t))$ 传送至栅格 A 类节点中，那么节点 i 于第一个时隙所发送的数据概率：$R_i(t)=(p_i^q(t)\times W_i^q(t_1,t_2))/\left(\sum_{i=1}^{n}p_i^q(t)\times W_i^q(t_1,t_2)\right)$。此时，节点 i 数据发送的概率有两种情况：

假设节点 i 于该刻竞争至发送时隙，那么 $R_i(t)=0$，栅格内剩下的节点在接下来的$[L_{data}^q/\mu_{real,i}(t)]$个时隙开启睡眠模式以节省能量，$i$ 传输数据完成之后，剩下的节点竞争信道；

假设节点 i 该时没竞争到信道，那么目前竞争到的时隙节点传输完数据之后，节点 i 将会重竞争信道，该时传输的概率为 $R_i(t)=R_i(t)+F\times(1-R_i(t))/n_{con}^i$，其中，$F$ 代表竞争的次数，n_{con}^i 代表节点 i 通信的范围内节点的数量，则缓存空间的阈值 $L_{\max}$ 的范围为：$(N*\pi*r^2)/(4*L^2)\leqslant n_{con}^i\leqslant(N*\pi*r^2)/L^2$。

综上，当周期 t 结束时，栅格内 A 类节点继续下一轮的平均速率、节点发送的优先级运算，利用这种更新新的节点得到信道概率的形式来缓解数据传输产生的冲突，降低拥塞率，提高数据传输完整度。

4 实验结果与分析

本文利用 Matlab 软件完成实验，节点的位置保持不变，

实验环境如下：将节点固定在 100 m×100 m 平面上，其中节点数据量为 60 个，实验场景如图 6 所示。运行 PC 机的配置：Pentium（R）4CPU2.40 GHz。

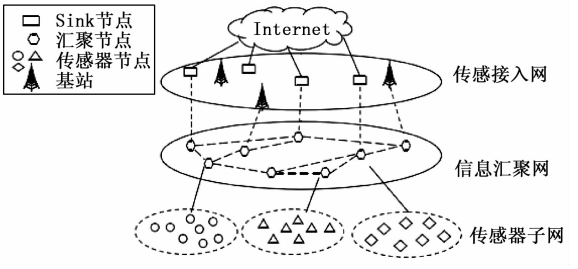

图 6　网络节点分布图

为了测试并行数据传输完整度控制方法的控制效果，将改进方法与传统方法对数据传输控制效果进行对比，如图 7 所示。

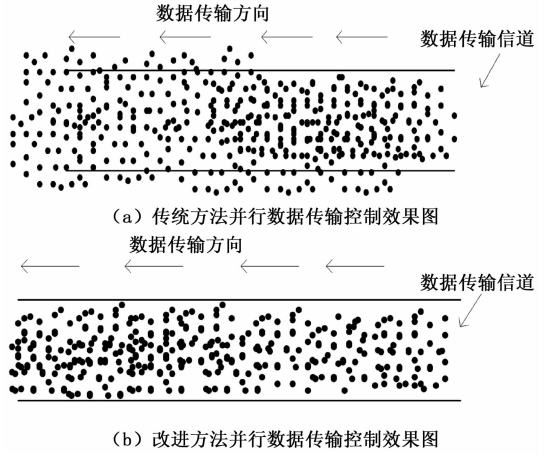

图 7　不同方法并行数据传输控制效果图

观察图 7（a）可知，采用传统方法对并行数据传输进行控制，数据在传输信道中轨迹十分散乱，有部分数据偏离了信道，导致数据传输过程中数据丢失。观察图 7（b）可知，采用改进方法对并行数据传输进行控制，数据均沿着信道传输，无任一数据偏离信道。对比图 7（a）和图 7（b）可得，改进方法比传统方法数据传输控制效果好，充分证明改进方法并行数据在网络传输过程中完整度高。

通过对比改进方法和传统方法并行数据传输节点的拥塞率，测试并行数据传输完整度控制方法的数据传输效率。两种方法并行数据传输节点拥塞率对比结果如图 8 所示。

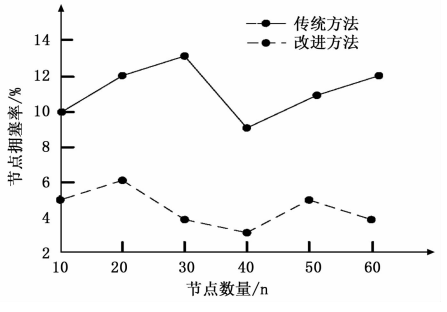

图 8　两种方法并行数据传输节点拥塞率对比图

根据图 8 可知，采用传统方法进行数据传输完整度控制，节点拥塞率平均保持在 11%左右，随着节点数量的增加，虽然节点数据为 40 个时，其拥塞率有下降情况，但总体数值呈上升趋势，且上升幅度较大。采用改进方法进行数据传输完整度控制，节点拥塞率平均保持在 4%左右，随着节点数量的增加，虽在节点数量为 50 个时，拥塞率有所上升，但总体数值呈下降趋势。对比两种方法可得，改进方法节点拥塞率远远低于传统方法的节点拥塞率，并持续降低，充分说明改进方法能够有效并行数据的传输速率，从而提高了并行数据传输的完整度。

为了测试并行数据传输完整度控制方法的控制精度，将传统方法与改进方法进行对比，两种方法并行数据传输完整度对比结果如图 9 所示。

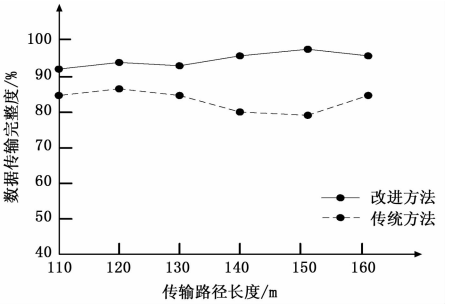

图 9　两种方法并行数据传输完整度对比图

观察图 9 可知，采用传统方法对并行数据传输完整度进行控制，传输后数据完整度平均为 83%，当传输路径长度为 150 m 时，数据完整度值最低，为 79%，且曲线波动较大。采用改进方法对并行数据传输完整度进行控制，其数据完整度平均为 95%，观察曲线，随传输路径长度的增加，总体呈平稳上升趋势。对比两种方法的实验曲线可得，改进方法的数据传输完整度高于传统方法的数据完整度，充分说明所提的并行数据传输完整度控制方法精度高，控制效果好。

分析上述实验结果可知，改进方法要绝对优于传统数据传输完整度控制方法。改进方法针对数据在信道传输中控制效果、节点拥塞率及数据完整度进行实验，通过对比两种方法的实验结果，验证所提并行数据传输完整度控制方法的优越性。该方法解决了节点级的拥塞现象，调整数据传输的速率，将数据在信道中传输的加以控制，减少数据偏离信道导致的丢失现象，保证了数据传输的完整性。

5　结束语

所提并行数据传输完整度控制方法的创新点在于对数据传输的单节点拥塞和系统级的阻塞进行了分析以及解决，这有效提高了数据传输的速率及完整度。未来会是一个互联网集成的环境，其中包含了异构式的技术和系统，由此在未来的研究中应对异构网络间的数据传输完整性方向做研究，以适应时代的发展。

参考文献：

[1] 蒋　俊，黄传河，华　超，等. 基于软件定义资源的实时控制 CPS 数据传输机制 [J]. 计算机工程与科学，2015，37（12）：2250－2255.

（下转第 188 页）