王宽诚教育基金会

学术讲座汇编

(第18集)

主编：钱伟长

上海大学出版社

王宽诚教育基金会《学术讲座汇编》

(第 18 集　1999 年)

钱伟长　主 编

上海大学出版社　上海市延长路 149 号
上海上工印刷厂印刷
开木 787 × 1092　1/16　彩插 1　印张 12 5　了数 295 千了
1999 年 12 月第 1 版　1999 年 12 月第 1 次印刷
印数 1—1050 册

ISBN 7-81058-151-1/Z ·004

目 录 CONTENTS

Molecular Dynamics Simulation and Stochastic Dynamics
Simulation in Biomolecular System ······ 施蕴瑜 (1)
THNPC- II -A Cluster Computing System with Fast NI Switch and
Message Passing Communication at User Level ······ 李三立 (7)
Memory Centric Interconnection Mechanism for Message
Passing in Parallel Systems ······ 李三立 (21)
Computational and Experimental Studies on the Thermolysis Mechanism
of Zirconium and Hafnium Tetraalkyl Complexes.
Difference between Titanium and Zirconium Complexes ······ 吴云东 (30)
谈大学生个人成长 ······ 岳晓东 (46)
中国文化和现代化管理 ······ 叶日崧 (53)
中国人口转变及其社会经济含义 ······ 朱国宏 (60)
人地关系与中国的社会经济变迁 ······ 朱国宏 (69)
近代中国大规模的国际迁移: 1840 ~ 1949 ······ 朱国宏 (82)
Synthesis and Characterization of Fe- and Cr-Containing Silicates Having Zsm-48
Structure Prepared with a Low Amount of 1,6-Hexanediamine ······ 李瑞丰 (87)
Catalytic Performances of Metallosilicates with Zsm-48 Structure ······ 李瑞丰 (92)
Catalytic Performances ofMetallosilicates with Zsm-48 Structure Preparation of
DLC Gradient Biomaterials by means of Plasma Source Ion Implant-Ion
Beam Enhanced Deposition ······ 尹光福 (97)
Magnetothermal Analysis of Fe-Cu-Nb-V-Si-B Alloy ······ 王 治 (103)
Temperature Dependence of Permeability for
$Fe_{72.7}Cu_1Nb_2V_{1.8}Si_{13.5}B_9$Nanocrystalline Alloy ······ 王 治 (108)
On the Auslander-reiten Quiver with Oriented Cycles
of Representation-finite Algebras ······ 姚海楼 (113)
Flow of Viscous Oil in a Centrifugal Oil Pump Impeller ······ 李文广 (117)
Method for Computing the Failure Probability
of Aircraft Structures ······ 宋笔锋 (126)
In Situ PCR and Immunohistochemical Studies
on p16 Genein Pituitary Adenomas ······ 易 静 (131)
Analysis with Non-Orthogonal Functions for Anisotropic Cells ······ 周哲玮 (139)
Nonlinear Vibration and Dynamic Instability of Thin Shallow Conical Shell ······ 叶志明 (147)
General Solution for C-Shaped Bellows Overall-Bending Problems ······ 朱卫平 (156)
Isotope Tracer Techniques in Cell Cultures for the Study of Plant Drugs in Relation
to receptors and post-receptor signal transduction ······ 胡雅儿 (163)
Labour Value and Water Resources Value ······ 甘筱青 (171)
On the Diesel Engine's Impulsive-source Features Extraction ······ 耿遵敏 (174)

王宽诚教育基金会
《学术讲座汇编》第18集，1～6 (1999)

Molecular Dynamics Simulation and Stochastic Dynamics Simulation in Biomolecular System

Yun-yu Shi[1]

School of Life Science, University of Science and Technology of China Laboratory of Structural Biology, University of Science and Technology of China, Chinese Academy of Science

1. Introduction

There are different methods used for study biomacromolecules, such as methods of biochemistry, molecular biology, cell biology, immunology ,X-ray diffraction, NMR, CD, ESR, IR, UV, fluorescence, Raman spectroscopy, EM, AFM, STM, and electron diffraction. Since 1970s, molecular modeling and simulation become another powerfull method. In the journal: *Current Opinion in Structural Biology*, theory and simulation become an important part of it.

The flowchart of simulation is shows in Fig.1.

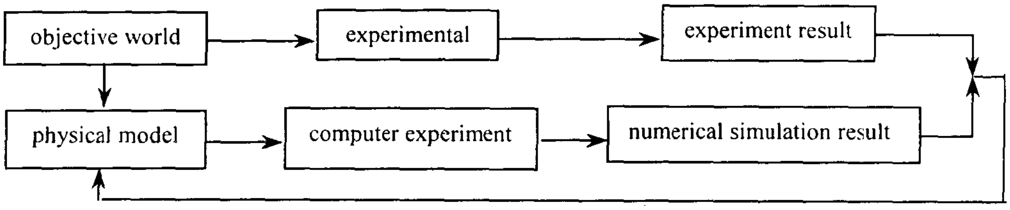


Fig.1 flowchart of simulation

A physical model was proposed first. Using computer to do numerical simulation, results were used to compare with experiments. Modify the physical model until results obtained coincide with experimental results. Then we can use this model to prediction other unknown properties.

Currently human genome projects are carrying on. After it, post-genome era will come. More and more people talk about the era bioinformatics will become more important.

Computer simulation can use to give explanation of the experimental results, to predict some properties that are difficult to be measured by experiments.

It may has following application:

(1) Three-dimensional structure prediction;

(2) Protein design, Properties prediction, such as stability, optimal pH, specificity Binding constant;

(3) De novo design of protein structure;

(4) Drug design. Many inhibitors of enzyme and antagonists of receptor are drugs;

(5) Simulation the conformational change and aggregation of protein.

[1] 施蕴瑜教授，中国科学院院士，中国科学技术大学结构生物学实验室主任（中国科学院院级开放实验室），由王宽诚教育基金会资助，于1998年5月赴香港科技大学，香港中文大学讲学，此为其讲稿。

In the theoretical research aspect:

(1) Studying biomolecular recognition Enzyme with substrate or inhibitor; antibody with antigen, protein with DNA or RNA, Protein with polysaccharides, protein with lipids;

(2) Modeling of enzyme catalyzed reaction mechanisms;

(3) Studding protein folding and stability;

(4) Studying electron transfer in photo synthesis reaction center;

(5) Studying ion transportation in ion channel.

2. Empirical force field

2.1 Atomic Potential Function

A typical molecular force field or effective potential for a biomacromolecular system of N atoms with mass m_i and Cartesian position vector r_i looks as following:

$$V = \sum_{i>j} V_{ij} + \sum_{I>j>k} V_{ik} + \sum_{i>j>K>l} V^{ijkl} + \cdots$$

$$V_i\ (\vec{r}) = V_i(\vec{r}_1, \vec{r}_2, \vec{r}_3, \cdots \vec{r}_N) =$$

$$\sum_{\text{bond}} \frac{1}{2} K_b (b - b_0)^2 + \sum_{\text{angle}} \frac{1}{2} K_\theta (\theta - \theta_0)^2 + \sum_{\text{tumable angle}} \frac{1}{2} K_\xi (\xi - \xi_0)^2 +$$

$$\sum_{\text{dihedral angle}} K_\varphi [1 + \cos(n\varphi - \delta)] + \sum_{ij} \left[\frac{C_{12}}{\gamma_{ij}^{12}} - \frac{C_6}{\gamma_{ij}^{6}} - \frac{q_i q_j}{4\pi \varepsilon_0 \varepsilon_\gamma r_{ij}} \right]$$

The first term represents the covalent bond stretching interaction along bond b_I. It is a harmonic potential in which the minimum energy bondlength b_0 and the force constant K_b vary with the particular type of bond. The second term describes the bond angle bending (three-body) interaction in similar form. Two forms are used for the dihedral angle interactions: a harmonic term for dihedral angles ξ that are not allowed to make transitions, e.g. dihedral angles within aromatic rings. and a sinusoidal term for the other dihedrals φ_i, who may make 360 degree turns. The last term is a sum over all pairs of atoms and represents the effective nonbonded interaction, composed of the Van der Waals' and the Coulomb interaction between atom i and atom j with charges q_i and q_j at distance r_{ij}.

2.2 Simplified Forms of Potential Function

United atom force field

Each amino acid is represented by one or two interaction point

Potentials of mean force from database

2.3 Interactions with Solvent

The water model use between 3~5 interaction sites, and a rigid water geometry. Following water models are general used.

TIP3P, SPC, SPC/E	3 sites
BP, TIP4P	4 sites
ST2	5 sites

3. Energy minimization

A system with N atoms, energy is a function of $3N$-6 internal coordinates or $3N$ Cartesian coordinates. According to the thermodynamic hypothesis that followed from Anfinsen's experiments on ribonuclease, a stable protein conformation corresponds to the lowest minimum of empirical conformational energy (including the effect of hydration)termed the global minimum. The multi-dimensional potential energy surface that describes the energy of polypeptide chain as a function of internal coordinates or Cartesian coordinates, however have an astronomically large number of local minima. It is not possible to explore the entire surface systematically in order to find the lowest minimum. Therefore, a major problem of conformational energy computations remains the efficient search of conformational space to locate the region the contains a potential well around the global energy minimum

The minimization problem can be formally stated as follows: given a function $f\,(x_1,x_2,x_3\ldots\ldots x_{3N})$, find the values of variables where f has a minimum value. At a minimum point the first derivative of the function with respect to each of the variables is zero and the second derivative are all-positive.

Minimization algorithms can be classified into two groups: those, which use derivative of the energy respective to the coordinates and those, which do not use derivative of the energy.

3.1 Non-derivative minimization methods

The simplex method

The sequential univariate method

3.2 Derivative minimization methods

(1) The steepest decent method

The steepest decent method moves in the direction parallel to the net force. For $3N$ Cartesian coordinates this direction is most conveniently represented by a $3N$ dimensional unit vector

$$S_k = -g_k/|g_k|$$

Where g_k is gradient. The new set of coordinates after step k is

$$X_{k+1} = X_k + \lambda_k S_k$$

λ_k is the step size

(2) Conjugate gradient method

In conjugate gradient method, the gradients at each point are orthogonal but the directions are conjugate. The conjugate gradient method moves a direction v_k from point X_k where v_k is computed from the gradient at the point and previous direction vector $v_{k\text{-}1}$

$$v_k = -g_k + \gamma_k v_{k\text{-}1}$$

γ_k is a scalar constant given by

$$\gamma_k = g_k g_k / g_{k\text{-}1} g_{k\text{-}1}$$

In the conjugated gradients method all of the directions and gradients satisfy the following relationship:

$$g_i \cdot g_j = 0$$
$$v_i \cdot V''_{ij} \cdot v_j = 0$$
$$g_i \cdot v_j = 0$$

4. Molecular Dynamics Simulation, MD

Stochastic Dynamics Simulation, SD

In the Molecular Dynamics Simulation, (MD)method a trajectory (configuration as a function of time)of the molecular system is generated by simultaneous integration of Newton's equations of motion

$$\frac{\mathrm{d}\,x_i}{\mathrm{d}\,t} = v_i$$

$$\frac{\mathrm{d}\,v_i}{\mathrm{d}\,t} = -\frac{1}{m}\frac{\partial V(\{x_i\})}{\partial x_i} i = 1,2,\ldots,N$$

for all the atoms in the system. MD simulation requires the calculation of gradient of potential energy $V(r)$. The integration is performed in small step Δt, typically 1-10 fsecs for molecular systems. Static equilibrium quantities can be obtained by averaging over the trajectory, which must be of sufficient length to form a representative ensemble of the state of system. In addition dynamical information can be extracted.

$$\langle Q\rangle = \overline{Q}(t) = \frac{1}{t}\int_0^t Q(r(t))\,\mathrm{d}\,t$$

$<V>$, $<E_{KIN}>$; $<r_i>$, $<\phi_i>$; $<(\triangle r)^2>^{1/2}$, $<(\triangle\ \phi)^2>^{1/2}$

$$C(t) = \langle \Delta\gamma(t)\Delta\gamma(t+\tau)\rangle$$

Stochastic Dynamics Simulation, SD

The method of stochastic dynamics (SD) is an extension of MD. A trajectory of the molecular system is generated by integration of the stochastic Langevin equation of motion

$$m_i\dot{v}_i(t) = F_i(\{x_i(t)\}) - m_i\gamma_i v_i(t) + R_i(t) \tag{1}$$

$F_i(\{x_i(t)\})$ is a mean force. $R_i(t)$.is a random force. Frictional forces proportional to a friction coefficient γ_i. The stochastic term introduces energy, the frictional term removes energy from the system.

$$< R_i(0)R_j(t) >= -2m_i\gamma_i KT_{ref}\delta_{ij}\delta(t)$$

$$\omega(R_i) = -[2\pi < R_i^2 >]^{1/2}\exp\{-R_i^2/(2 < R_i^2 >)\}$$

$$< R_i >= 0$$

$$< v_i(0)R_i(t) >= 0 \qquad t \geq 0$$

$$< F_i(0)R_i(t) >= 0 \qquad t \geq 0$$

$$\gamma_i = (6\pi R\eta / M)\omega_i \tag{2}$$

SD can be applied to establish a coupling of individual atom motion to a heat bath, or to mimic solvent effect. In the latter case the stochastic term represents collisions of solute atoms with solvent molecules and the frictional term represents the drag exerted by the solvent on the solute atom motion.

Generalized Langevin dynamics (GLD) simulation

$$m_i\ddot{x}_i(t) = F_i(t) - m_i\int_0^t\gamma_i(t-\tau)\dot{x}_i(\tau)\,\mathrm{d}\,\tau + R_i(t) \tag{3}$$

$$\langle R_i(0)R_i(t)\rangle = m_i k_B T\gamma_i(t) \tag{4}$$

The solvent effect can be divided as

(1) The average interaction between solute atoms is affected by the presence of solvent.

(2) A dynamic effect on the solute, a frictional force solvent, and a randomly fluctuating force representing collision with solvent molecules

The average effects of solvent atoms

$$V_{\text{mean}}(\{x_i\}) = \langle V(\{x_i\}, \{x_\alpha\}) \rangle_\alpha \qquad (5)$$

A process of transferring a molecule from vacuum to solvent the solvent influence on solute can be considered to have three components

$$G_{\text{sol}}(\{x_i\}) = G_{\text{ele}}(\{x_i\}) + G_{\text{disp}}(\{x_i\}) + G_{\text{cav}}(\{x_i\})$$

There are three type of approaches describing electrostatic effect of solvent: (1) empirical methods related to solvent accessible surface of a solute. (2) microscopic methods. (3) methods based on the continuum electrostatics model. A rigorous and computational feasible approach to the treatment of electrostatic interaction of solvent has been provided by development of algorithms for the numerical solution of Poisson-Boltzmann (PB) equation of classical electrostatics.

Advantages of PB based approach are:

(1) Reduction of size of system

(2) Accessible of time scale of the simulation

Poisson equation and Poisson-Boltzmann equation are:

$$\nabla^2 \phi_p^I = \frac{1}{D_i} \sum_k q_k \delta(\bar{r}_k - \bar{r}_p) \qquad (6)$$

$$\nabla^2 \phi_p^E = \kappa^2 \phi^E(\bar{r}_p) \qquad (7)$$

Boundary conditions:

1) $\phi^I = \phi^E$

2) $\varepsilon_i \dfrac{\partial \phi^I}{\partial n} = \varepsilon_e \dfrac{\partial \phi^E}{\partial n}$

Numerical methods are frequently used to solve the PBE:

(1) Finite difference.method (FDM)

(2) Finite element method (FEM)

(3) Boundary element method (BEM)

We have developed the method for calculate solvation energy with Boundary Element Method(BEM) . Then use this method study pKa of active group in protein and titration curve of protein.

In the BEM method, using following definitions:

$$\lambda_{pq} = \frac{1}{4\pi r_{pq}}$$

$$\mu_{pq} = \frac{\exp(-Kr_{pq})}{4\pi r_{pq}}$$

Where r_{pq} is the distance between point p and q, K is a constant.

The interior potential f^I and exterior potential f^E are governed by Eqs.(8~9) respectively:

$$\Delta^2 \phi_p^I = \frac{1}{D_i} \sum_k q_k \delta(r_k - r_p) \tag{8}$$

$$\Delta^2 \phi^E (r_p) = K^2 \phi^E (r_p) \tag{9}$$

r_p is a point inside or outside the molecule,

r_k is the position the kth charge

K is Debye inverse screening constant.

Using Green's second identity on Eqs.(8~9), we obtain following boundary integral equations for the interior and exterior potentials correspondingly:

$$\frac{1}{2}\phi_i^I = G_i + \int_s \frac{\partial \phi_j^I}{\partial n} \lambda_{ij} \, \mathrm{d}\, A_j - \int_s \phi_j^I \frac{\partial}{\partial n}(\lambda_{ij}) \, \mathrm{d}\, A_j \tag{10}$$

$$\frac{1}{2}\phi_i^E = -\int_s \frac{\partial \phi_j^E}{\partial n} \mu_{ij} \, \mathrm{d}\, A_j + \int_s \phi_j^E \frac{\partial}{\partial n}(\mu_{ij}) \, \mathrm{d}\, A_j \tag{11}$$

Where i and j are the points on the surface. The surface potentials and their normal derivatives satisfy following boundary conditions:

$$\phi^I = \phi^E \tag{12}$$

$$D_i \frac{\partial \phi^I}{\partial n} = D_e \frac{\partial \phi^E}{\partial n} \tag{13}$$

We have incorporating hydration force determined by boundary element method into stochastic dynamics. Then use this method to study conformation and dynamical properties of cyclosporin A and alanin dipeptide.

References

1 Xiang Z.X., Shi Y.Y., Xu Y.W. Calculating the electric potential of macromolecules: A simple method for molecular surface triangulation. *J. Comput. Chem.*, 1995; **16**: 512-516

2 Xiang Z.X., Huang F.H., Shi Y.Y. Calculation of solvation energy with a conbination of boundary elementary method and PDLD model. *J. Phys. Chem.*, 1994; **98**: 12782-12788

3 Wang C.X., Wan S.Z., Xiang Z.X., Shi Y.Y. Incorperating hydration force determined by boundary elementary meythod into stochastic dynamics, *J. Phys. Chem.B*, 1997; **101**: 230-235

4 Wan S.Z., Wang C.X., Xiang Z.X., Shi Y.Y. Stochastics Simulation of the Ala Dipeptide: Including solvation interaction determined by boundary elementary method, *J. Comp.Chemistry*, 1997; **18**(12): 1440

5 Wan S.Z., Wang C.X., Shi Y.Y. Generalized Langevin dynamics simulation: numerical integration and application of the generalization Langevin equation with an exponential model for the friction kernel. *Molecular Physics*, 1998; **93**(6): 901-912

王宽诚教育基金会
《学术讲座汇编》第 18 集，7～20 (1999)

THNPC-Ⅱ a Cluster Computing System with Fast NI Switch and Message Passing Communication at User Level

Prof. Li San-li[1]

Abstract

THNPC Ⅱ is a cluster computing system combining Networked Parallel Computing Environment with Interconnection Network (IN) technique of MPP. Network bandwidth and software overhead constitute two essential factors influencing on performance of cluster computing systems. Replacing the conventional common medium network, the IN switch of THNPC Ⅱ employs the custom designed fast switch with circulating priority scheduling, cut-through and data-partitioning techniques similar to the FLIT of Worm-hole Routing used in MPP. It provides a improved network bandwidth in comparison with the conventional interconnection switches used in cluster computing. For the purpose of reducing the software overhead in message passing systems, we design a high speed Network Interface Adaptor (NIA) with PCI bus interface. The applications access the IN switch via this NIA to carry out the message passing communication at user level, so that it can decrease the overhead of system driver program and traditional network protocol. Furthermore, it allows the message passing communication between computing nodes to be overlapped with computation. Experimental measurement is depicted, and the result is given.

1. Introduction

In recent years, Cluster Computing (sometimes called Networked Parallel Computing, NPC) has become a promising approach to achieving supercomputing due to its friendly user environment and flexible construction. Software overhead and interconnection network (IN) bandwidth constitute two essential factors affecting on the performance of cluster computing systems. Recently, it emerges a tendency in combing the IN technique of MPP with cluster computing environment, ordinarily the cluster computing systems employ traditional common medium Ethernet or Ethernet switch. The IN of MPP provides higher bandwidth and lower communication latency than that of conventional Ethernet, this kind of fast IN switch can connect a number of computing nodes, which are commercially available PC/Workstation Systems within Moderate Communication Distance to achieve high peak

[1] 李三立教授，中国工程院院士，上海大学计算机学院院长，清华大学计算机科学与工程研究所所长，由王宽诚教育基金会赞助，于 1998 年 12 月赴香港大学讲学，此为其讲稿。

performance, we call this kind of cluster computing system NPC/MOCOD. However, NPC/MOCOD carries out the communication between computing nodes by message passing via conventional LAV interconnection network protocol, that causes a large portion of software overhead in parallel computing.

This paper describes a cluster computing system at Tsinghua University termed THNPC-II, in which each computing node employs low cost twin Pentium-II SMP system, shown in Fig.1. It employs TH-Net as its interconnection network built on the basis of routers which has the features of circulating priority scheduling, cut through and data partitioning, similar to the FLIT of worm hole routing technique used in many MPPs. This THNPC-II system also adopted the custom designed Network Interface Adaptor (NIA) with PCI bus interface, in order to match the high bandwidth of IN. The NIA implements the message-passing communication at user level, thus it greatly reduces the software overhead resulted by traditional network protocol and system driver program, the NIA design also allows the communication to be overlapped with useful computation in computing nodes, so that it furthermore improves the performance of THNPC-II cluster computing system.

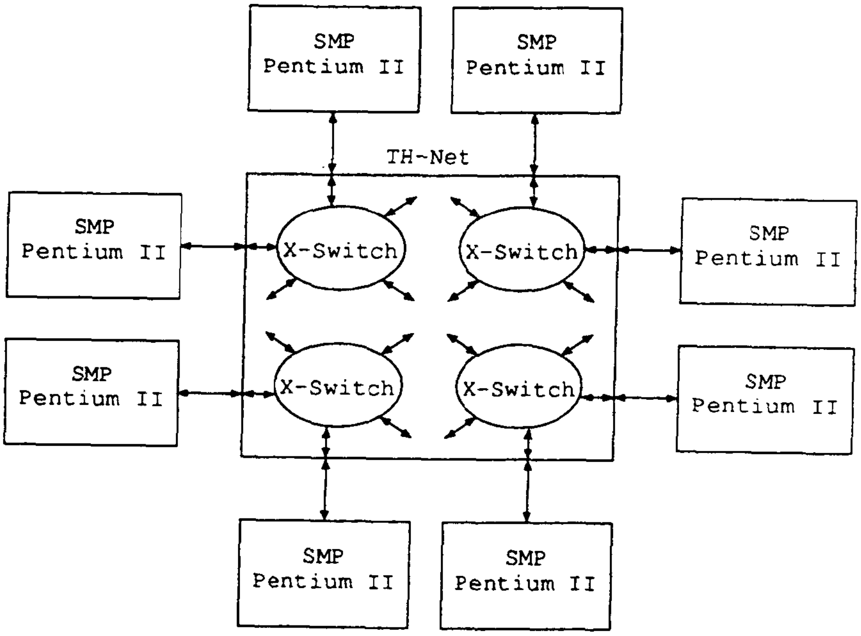


Fig.1 THNPC-II system with its TH-Net

This paper also describes the experiment of a case study evaluation and gives the experimental performance evaluation in the last section.

2. Description of TH-Net Structure

2.1 X-Switch of TH-Net

The TH-Net is built on the basis of router, called X-Switch, each X-Switch has five pairs of Input/Output ports for linking with the other X-Switch or NIA of nodes. As shown in Fig.2a, one port connects with the neighbor port or the NIA interface of the opposite side to construct on X-Link. X-link provides a pairs of bidirectional x-channels between two connected sides for linking. As shown in Fig. 2 b & c, by using X-links and X-switches it is easy and flexible to construct the cluster computing systems with different topologies and good scalability.

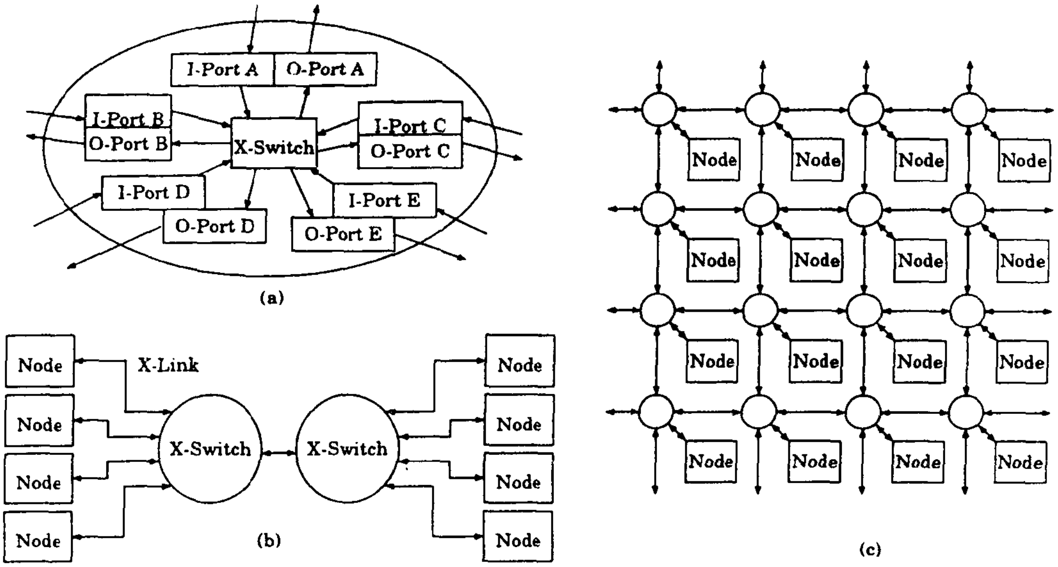


Fig.2 X-Switch with five pairs of ports (a), and different topologies of interconnection for Cluster Computing Systems (b),(c)

2.2 Packet format and routing control

The width of data unit in X-link is 9 bits. The 9-bit data stream forms a Data Frame which contains a Frame Head, Valid Data, a Frame End Tag and some check bits, the Data Frame format is shown in shown in Fig.3. Beyond the conventional data, there are control characters using the same data width to be transmitted. A "O" in the most significant bit indicates conventional data, differentiated from a control word with a highest bit"1". The control characters include the signals of suspending data transmission (ID-STOP), resuming data transmission (ID-GO), filling the timing vacancy slot (ID-IDLE), and ending a Data Frame (ID-EOF), as well as the routing characters, ID-FA/B/C/D/E, indicating the transmission of a Data Frame to output port of and X-Switch, where A,B,C,D and E represent five ports of and X-Switch separately.

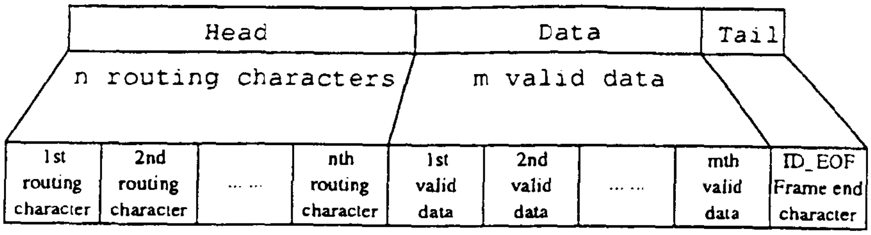


Fig.3 Data frame format

Path selection in TH-Net is determined by the sending node. In order to enable dynamic routing control by hardware for reducing the latency, the destination of a Data Frame is not represented by the destination address, instead, by the relative path. For instance, in TH-Net, node P intends to transmit data to another node Q. Suppose that data would pass through 3 X-Switches. In this situation, the destination of the Data Frame is specified by a path represented by (IN-FB, ID-FE, ID-FC). Since the topology of the network is statically determinable, each node saves the path needed for the current topology, and, the sending node determines a transmission path, shown by the sequence of several switch port numbers in the Head of a Data Frame. Both of the Head and Content of a Data Frame are variable in length. Therefore, all the routing characters in the Head of a Data Frame are used to designate a path of transmission through TH-Net. As the X-Switches pass this Data Frame, it eliminates one routing character from the Head at each switch. So, when all the routing characters in the Head have been eliminated, this Data Frame has reached its destination.

2.3 Flow control and scheduling strategy in X-Switch

As we mentioned above, an X-link is composed of a pair of bidirectional X-channels. The channel transmits synchronized data, i.e. signals are sent and received with a fixed clock frequency. Each In-Port has a FIFO Buffer, when the data receiving side notices that the Buffer is nearly full, it requests the data sending side to suspend data transmission temporarily. When the sending side receives a suspend request, it will stop transmission and enter into a waiting state (transmitting the ID-IDLE idle character on the channel), until the data in the buffer of the receiving side have been processed (data extracted from the buffer), then the data receiving side sends the resume data transmission command (ID-GO)to the sending side. Afterwards the sending side transits from the waiting state into the data transmission state and continues to transmit data. The flow control commands, such as ID-STOP/ID-GO, are inserted by receiving side into the data stream of the reverse direction of X-channel. The flow control request must not be delayed, so the insertion of flow control commands has the highest priority level in IN-Net, in order to ensure that flow cntrol commands are sent immediately.

If Data Frames coming from more than one Input Port of an X-Switch request the same Output Port, a block would occur. Blockade in TH-Net is caused by the contention for Output Port resource. In our design, an idle Output Port adopts a circulating priority strategy for serving Input Ports. The five ports, A,B,C,D and E, obtain the access right to the same Output Port in a round-robin mode. Using this feature, we can divide a larger Data Frame into several small Data Frames (just like FLITs in the Worm hole routing[4]). If three Input Ports, A, B and C in one X-switch of TH-Net, have their data reaching simultaneously at the same Output Port D, but the Data Frame coming from Port A is divided into smaller frames A_0,A_1,A_2···An, similarly the Data Frames from Port B are B_0,B_1,B_2,···and from Port C are C_0,C_1,C_2···In this case, Port D allocates the priority to Port A, B and C in a round-robin mode for each smaller Data Frame. The data stream output from Port D will be A_0, B_0, C_0, A_1, B_1, C , A_2, B_2, C_2,···

By adopting the circulating priority scheduling and designing and Input Buffer in an X-Switch, in addition to the automatic division of larger Data Frames into smaller Data Frames, it can achieve good results, similar to using Virtual Channel Technology(VCT)[11].

Furthermore, each Input Port only needs its FIFO butter, requiring simpler design than VCT. The combination of the above mentioned techniques usually used in MPPs, can prevent frequent blockages, and can provide high scalability in cluster computing. The Flow-Control-Circuit in Input and Output Ports of X-Switch is shown in Fig.4. If it is used in a massive mesh topology structure, a high effective bandwidth for the interconnection network can be assured in NPC/MOCOD. It resolves the problems of low network bandwidth and low scalability introduced by a conventional network, meanwhile, from the hardware point view, it ensures the delay of message passing to be greatly reduced. The implementation of flow control can be illustrated by a state machine.

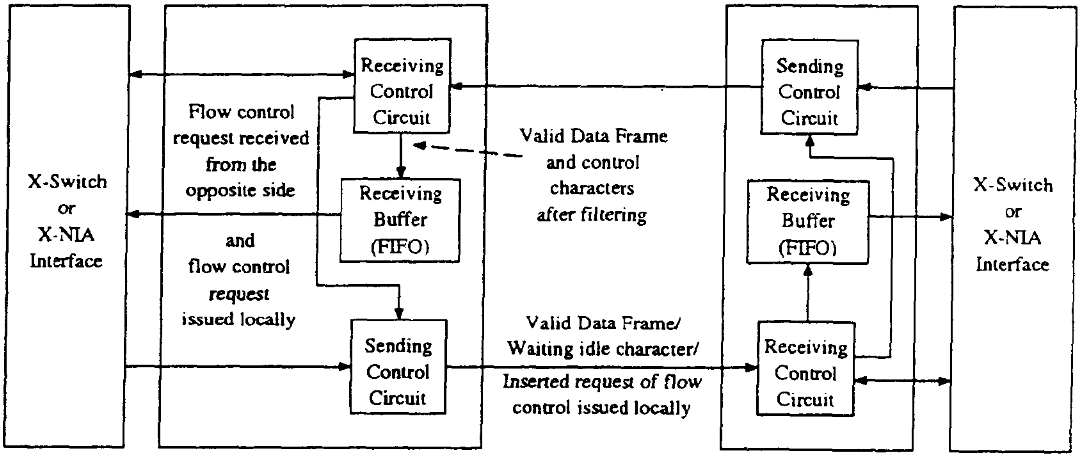

X-Switch or X-NIA Interface
Flow control request received from the opposite side
and flow control request issued locally
Receiving Control Circuit
Receiving Buffer (FIFO)
Sending Control Circuit
Valid Data Frame and control characters after filtering
Valid Data Frame/ Waiting idle character/ Inserted request of flow control issued locally
Sending Control Circuit
Receiving Buffer (FIFO)
Receiving Control Circuit
X-Switch or X-NIA Interface

Fig.4 Structure of the flow control circuit in input port and output port

The state machine for Output port is shown in Fig.5. This state machine automatically selects the Output Port A according to the routing control characters in the Head of the Data Frame, automatically analyses the control words from the dual directional data streams and implements the flow control strategy. Each Output Port has a state machine control mechanism. The state machine has 15 states. When the Output Port is not in use, the output state machine automatically polls each Input Port (states Poll A~Poll E) to detect whether any port wants to transmit a Data Frame to its Output Port. If it receives a Data Frame request, the state machine automatically transits into a transmission state (TranA2A~TranE2A).

In the states TranA2A~TranE2A, data is being transmitted from Input ports A~E to the corresponding local Output Ports. In the transmission state, if a Frame End (IN-EOF) is detected, then the state machine automatically breaks off the data link and transits into a polling state. If it is in the transmission state and it receives flow control commands, then the state machine suspends the data transmission temporarily, reverse the data link, then transits into wait states, Wait A2A~Wait E2A.

In the wait states, the data link from Input Port A~E is reversed, only flow control commands can be transmitted through Output Ports: the transmission of data and the other control characters must be prohibited. Idle (wait) cycles will be filled with ID-IDLE characters on the data link. The link will remain in the wait states until the Output Port will receive an ID-GO signal generated by the opposite side of communication, then it will resume transmission by transitting to the corresponding transmission state, TranA2a~TranE2A.

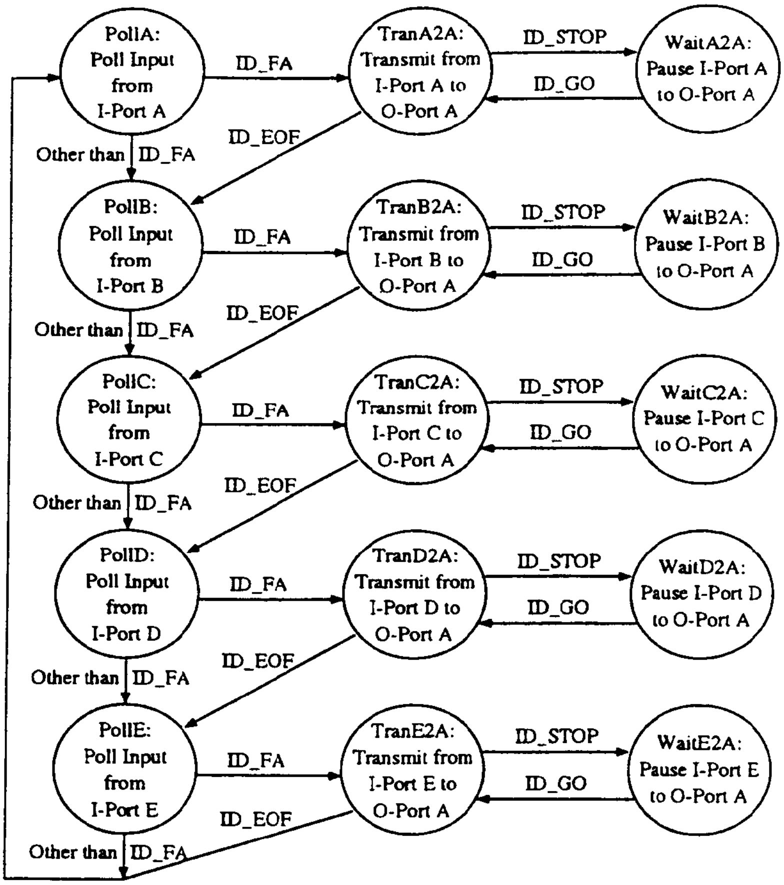


Fig.5 The scheduling state machine

3. Network Interface Adaptor Supporting Message Passing at User Level

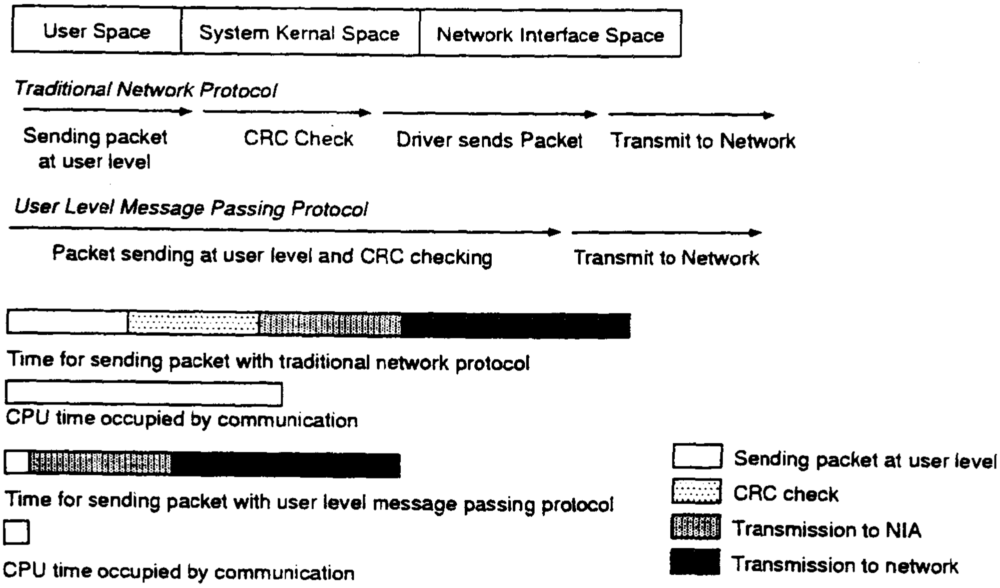
The design of network Interface Adapter (NIA)in cluster computing system should match the high performance of Interconnection Network. In a message-passing based networked parallel computing system, the transmission of a massage usually requires passing through multiple levels of network protocol and switching across the boundary between kernel and user level many times. It constitutes considerably high software overhead in networked parallel computing systems. Many researchers[1,2,6] have studied that carrying out the communication mechanism at use level is an important approach to reducing the software overhead.

We have designed a NIA, called X-NIA, or THNPC-II, which supports the message passing communication at user level, and allows the communication in the IN to be overlapped with the useful computation in computing nodes.

3.1 Message passing at user level

The central idea of message passing at user level is to provide the message passing function interface of directly accessing the NIA for applications, and during the invocation of these functions the system keeps running at the user level. It is also needed to shorten the function invocation time.

As shown in Fig.6, in the communication procedure of traditional network protocol, the software overhead is caused by status switching many times during the system function invocation, the data packets have been copied at least two to three times in system memory, and the time for each data copy is almost equal to the message transmission time in the interconnection network, and these operations occupy the CPU resource of this system. CPU not only is required to carry out data copy, but also to generate CRC check. While in the message passing at user level, it needs data copy only once, meantime, usually it reads the data packet directly from system memory by using DMA mode designed in the network interface adaptor NIA. Thus it doesn't occupy CPU time. Therefore, the message passing at user level can significantly reduce the software overhead.



User Space
System Kernal Space
Network Interface Space
Traditional Network Protocol
Sending packet at user level
CRC Check
Driver sends Packet
Transmit to Network
User Level Message Passing Protocol
Packet sending at user level and CRC checking
Transmit to Network
Time for sending packet with traditional network protocol
CPU time occupied by communication
Time for sending packet with user level message passing protocol
CPU time occupied by communication
Sending packet at user level
CRC check
Transmission to NIA
Transmission to network

Fig.6 Data packet transmission time analysis

3.2 Design consideration

For the purpose of conducting message passing at user level, it should allow to directly access the NIA at user level, i.e. to issue the command of sending message directly to NIA; Besides, it requires that NIA can directly access the memory.

There are many kinds of accessing NIA. The earlier method is to set up I/O access address for NIA, and then by using IN/OUT instructions to access the registers on NIA, to issue command and to poll the status. However, the more efficient way is to take address mapping for the register/ memory of NIA, the applications can access NIA with memory reference instructions, but just like I/O ports, in the concrete design, it is still needed to protect the critical resources.

Recently, Eicken *et al.*[12] employ the DMA support by network interface, modify the system network driver programs, and provide the function libraries supporting message passing at user lever. SHRIMP implements the network interface that can directly access system memory[15], and then constructs a distributed shared memory on this basis to allow the applications to exchange data at user level.

In our design, we make use of PCI bus specification that supports the interface card to directly access the system memory, meanwhile to keep the memory Cache coherency maintained by system. In such case, if the applications at user level can directly inform the NIA of the data transmission request, then NIA can actively access the data in main memory with avoiding the interference from system kernel. Similarly, during date receiving, it also allows the NIA to directly write data into main memory.

3.3 X-NIA hardware structure

We have designed the NIA for THNPC-II, called X-NIA. The structure of X-NIA card is shown in Fig.7. It consists of Control Processing Unit (ICU), Register Files, Static Memory and Peripheral Interface Unit.

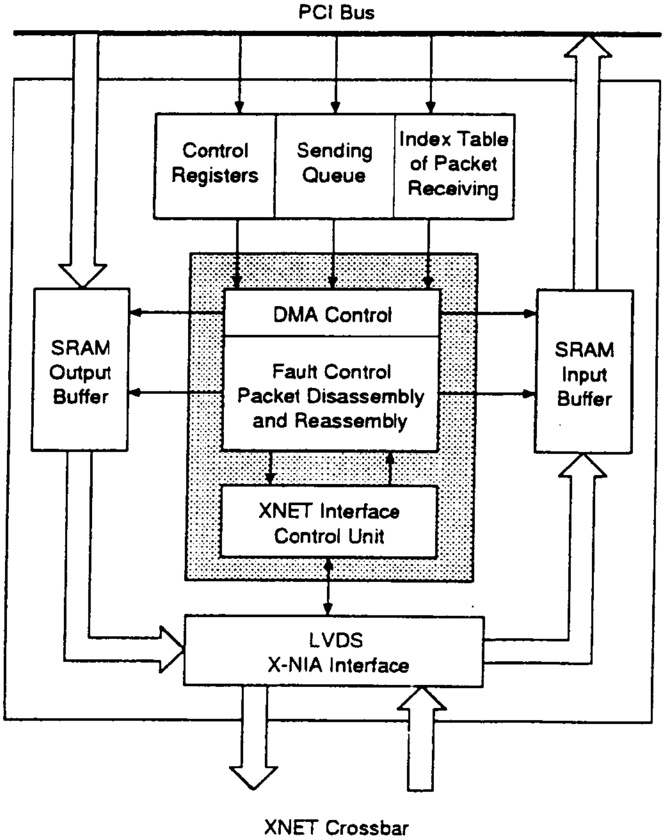


Fig.7 Structure of X-NIA card

Register files can be divided into three kinds: Control Registers, Sending Queue and Index List of