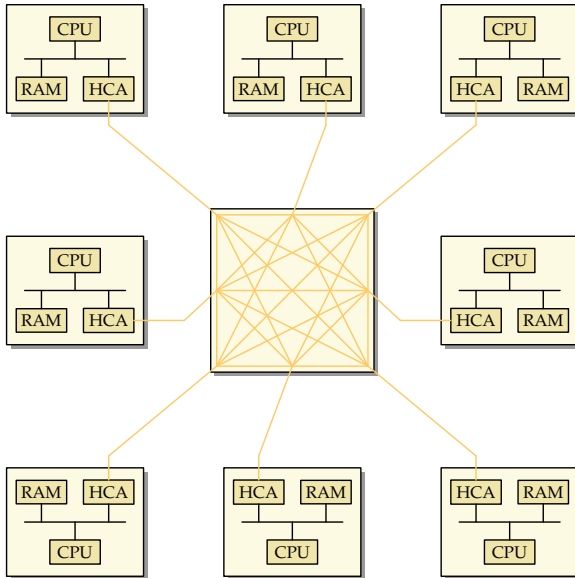


- Multicomputer bestehen aus einzelnen Rechnern mit eigenem Speicher, die über ein Netzwerk miteinander verbunden sind.
- Ein direkter Zugriff auf fremden Speicher ist nicht möglich.
- Die Kommunikation kann daher nicht über gemeinsame Speicherbereiche erfolgen. Stattdessen geschieht dies durch den Austausch von Daten über das Netzwerk.

- Eine traditionelle Vernetzung einzelner unabhängiger Maschinen über Ethernet und der Verwendung von TCP/IP-Sockets erscheint naheliegend.
- Der Vorteil ist die kostengünstige Realisierung, da die bereits vorhandene Infrastruktur genutzt wird und zahlreiche Ressourcen zeitweise ungenutzt sind (wie etwa Pools mit Desktop-Maschinen).
- Zu den Nachteilen gehört
  - ▶ die hohe Latenzzeit (ca.  $150\mu s$  bei GbE auf PacioLi, ca.  $500\mu s$  über das Uni-Netzwerk),
  - ▶ die vergleichsweise niedrige Bandbreite,
  - ▶ das Fehlen einer garantierten Bandbreite und
  - ▶ die Fehleranfälligkeit (wird von TCP/IP automatisch korrigiert, kostet aber Zeit).
  - ▶ Ferner fehlt die Skalierbarkeit, wenn nicht erheblich mehr in die Netzwerkinfrastruktur investiert wird.

- Mehrere Hersteller schlossen sich 1999 zusammen, um gemeinsam einen Standard zu entwickeln für Netzwerke mit höheren Bandbreiten und niedrigeren Latenzzeiten.
- Infiniband ist heute die populärste Vernetzung bei Supercomputern: Vier Supercomputer der TOP-10 und 213 der TOP-500 verwenden Infiniband (Stand: November 2010).
- Die Latenzzeiten liegen im Bereich von 140 *ns* bis 2,6  $\mu s$ .
- Brutto-Bandbreiten sind zur Zeit bis ca. 40 Gb/s möglich. (Bei Pacioli: brutto 2 Gb/s, netto mit MPI knapp 1 Gb/s.)
- Nachteile:
  - ▶ Keine hierarchischen Netzwerkstrukturen und damit eine Begrenzung der maximalen Rechnerzahl,
  - ▶ alles muss räumlich sehr eng zueinander stehen,
  - ▶ sehr hohe Kosten insbesondere dann, wenn viele Rechner auf diese Weise zu verbinden sind.

- Bei einer Vernetzung über Infiniband gibt es einen zentralen Switch, an dem alle beteiligten Rechner angeschlossen sind.
- Jede der Rechner benötigt einen speziellen HCA (*Host Channel Adapter*), der direkten Zugang zum Hauptspeicher besitzt.
- Zwischen den HCAs und dem Switch wird normalerweise Kupfer verwendet. Die maximale Länge beträgt hier 14 Meter. Mit optischen Kabeln und entsprechenden Adaptern können auch Längen bis zu ca. 100 Meter erreicht werden.
- Zwischen einem Rechner und dem Switch können auch mehrere Verbindungen bestehen zur Erhöhung der Bandbreite.
- Die zur Zeit auf dem Markt angebotenen InfiniBand-Switches bieten zwischen 8 und 864 Ports.



Die extrem niedrigen Latenzzeiten werden bei InfiniBand nur durch spezielle Techniken erreicht:

- ▶ Die HCAs haben direkten Zugang zum Hauptspeicher, d.h. ohne Intervention des Betriebssystems kann der Speicher ausgelesen oder beschrieben werden. Die HCAs können dabei auch selbständig virtuelle in physische Adressen umwandeln.
- ▶ Es findet kein Routing statt. Der Switch hat eine separate Verbindungsleitung für jede beliebige Anschlusskombination. Damit steht in jedem Falle die volle Bandbreite ungeteilt zur Verfügung. Die Latenzzeiten innerhalb eines Switch-Chips können bei 200 Nanosekunden liegen, von Port zu Port werden beim 648-Port-Switch von Voltaire Latenzzeiten von 100-300 Nanosekunden erreicht.

Auf PacifiCl werden auf Programmebene (mit MPI) Latenzzeiten von unter 5  $\mu s$  erreicht.

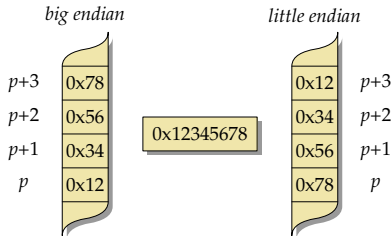
- Mit einer Rechenleistung von 2,5 Petaflop/s ist die Installation seit Oktober 2010 der leistungstärkste Rechner.
- Zum Einsatz kommen 14.336 Xeon X5670-Prozessoren und 7.168 Nvidia Tesla M2050-GPUs.
- Ein einzelner Knoten verfügt über 2 Xeon-Prozessoren und einer GPU. Jeweils zwei Knoten werden zu einer Blade zusammengefasst.
- 8 Blades und ein Switching-Board mit 16 Ports bilden eine Einheit. Vier solcher Einheiten werden in einem Schrank zusammengefasst. Insgesamt gibt es 112 solcher Schränke.
- Die Kommunikation erfolgt über ein proprietäres Hochgeschwindigkeitsnetzwerk mit 160 Gb/s.

- Mit einer Rechenleistung von mehr als einem Petaflop/s war es bis Juni 2009 die leistungsstärkste Rechnerinstallation, inzwischen nur noch auf Platz 7.
- Das System hat eine hybride Architektur mit 12.960 IBM PowerXCell 8i CPUs und 6.480 AMD-Opteron-Dual-Core-Prozessoren.
- Ein Triblade besteht aus zwei AMD-Opteron-Prozessoren mit 16 GB Hauptspeicher und vier PowerXCell 8i CPUs, ebenfalls mit 16 GB Hauptspeicher.
- Eine CU (Connected Unit) besteht aus 180 TriBlades, die sich über 15 Racks verteilen. Zu jeder CU gehört ein 288-Port-Switch (Voltaire ISR 2012).
- Insgesamt gibt es 18 CUs.
- Mehrere 288-Port-Switches von Voltaire lassen sich zu einem virtuellen Gesamtswitch kombinieren, so dass eine Zweistufenhierarchie entsteht.
- Der RoadRunner hat 8 288-Port-Switches auf der zweiten Ebene. Jeder der CU-Switches hat je 12 Verbindungen zu jedem der 8 Switches der zweiten Stufe.



- 36 Knoten mit jeweils 2 AMD-Opteron-Prozessoren (2,6 GHz).
- Jeder der Knoten verfügt über 8 GiB Hauptspeicher und eine Festplatte.
- Über einen Switch sind sämtliche Knoten untereinander über Infiniband vernetzt.
- Zusätzlich steht noch ein GbE-Netzwerk zur Verfügung, an das jeder der Knoten angeschlossen ist.
- Pacioli bietet zwar auch IP-Schnittstellen für Infiniband an, aber diese sind nicht annähernd so effizient wie die direkte Schnittstelle.

- Da einzelne Rechner unterschiedlichen Architekturen angehören können, werden möglicherweise einige Datentypen (etwa ganze Zahlen oder Gleitkommazahlen) unterschiedlich binär repräsentiert.
- Wenn die Daten mit Typinformationen versehen werden, dann wird die Gefahr von Fehlinterpretationen vermieden.
- Die Übertragung von Daten gibt auch die Gelegenheit, die Struktur umzuorganisieren. Beispielsweise kann ein Spaltenvektor in einen Zeilenvektor konvertiert werden.
- Auch die Übertragung dynamischer Datenstrukturen ist möglich. Dann müssen Zeiger in Referenzen umgesetzt werden.
- Die Technik des Verpackens und Auspackens von Datenstrukturen in Byte-Sequenzen, die sich übertragen lassen, wird Serialisierung oder *marshalling* genannt.



- Bei *little endian* sind im ersten Byte die niedrigstwertigen Bits (hier 0x78).
- Die Reihenfolge ist bei *big endian* genau umgekehrt, d.h. die höchstwertigen Bits kommen zuerst (hier 0x12).
- Da der Zugriff immer byte-weise erfolgt, interessiert uns nur die Reihenfolge der Bytes, nicht der Bits.
- Zu den Plattformen mit *little endian* gehört die x86-Architektur von Intel, während die SPARC-Architektur normalerweise mit *big endian* operiert.

- Prinzipiell hat sich als Repräsentierung das Zweier-Komplement durchgesetzt:

$$a = \sum_{i=1}^{n-1} a_i 2^{i-1} - a_n 2^n$$

- Wertebereich:  $[-2^{n-1}, 2^{n-1} - 1]$
- Dann bleibt nur noch die Entscheidung über die Größe von  $n$  und die Reihenfolge der einzelnen Bytes.
- Bei letzterem wird traditionell *big endian* gewählt (*network byte order*), siehe RFC 791, *Appendix B*, und RFC 951, 3. Abschnitt.

- Durch die *Google Protocol Buffers* wurde eine alternative Repräsentierung populär, bei der ganze Zahlen mit einer variablen Anzahl von Bytes dargestellt werden.
- Von den acht Bits wird das höchstwertige nur als Hinweis verwendet, ob die Folge fortgesetzt wird oder nicht: 1 = Folge wird fortgesetzt, 0 = Folge ist mit diesem Byte beendet.
- Die anderen sieben Bits (deswegen zur Basis 128) werden als Inhalt genommen, wobei sich Google für *little endian* entschied, d.h. die niedrigstwertigen Bits kommen zuerst.
- Dieses Verfahren ist jedoch ungünstig für negative Zahlen im Zweierkomplement, da dann für die -1 die maximale Länge zur Kodierung verwendet werden muss.
- Beispiel: 300 wird kodiert als Folge der beiden Bytes 0xac und 0x02 (binär: 1010 1100 0000 0010).

- Bei vorzeichenbehafteten ganzen Zahlen verwenden die *Google Protocol Buffers* die sogenannte Zickzack-Kodierung, die jeder ganzen Zahl eine nicht-negative Zahl zuordnet:

ganze Zahl	Zickzack-Kodierung
0	0
-1	1
1	2
-2	3
2147483647	4294967294
-2147483648	4294967295

- Das bedeutet, dass das höchst- und das niedrigstwertige Bit jeweils vertauscht worden sind. Bei 32-Bit-Zahlen sieht das dann so aus:

$$(n \ll 1) \wedge (n \gg 31)$$

- IEEE-754 (auch IEC 60559 genannt) hat sich als Standard für die Repräsentierung von Gleitkommazahlen durchgesetzt.
- Eine Gleitkommazahl nach IEEE-754 besteht aus drei Komponenten:
  - ▶ dem Vorzeichen  $s$  (ein Bit),
  - ▶ dem aus  $q$  Bits bestehenden Exponenten  $\{e_i\}_{i=1}^q$ ,
  - ▶ und der aus  $p$  Bits bestehenden Mantisse  $\{m_i\}_{i=1}^p$ .
- Für **float** und **double** ist die Konfiguration durch den Standard festgelegt, bei **long double** ist keine Portabilität gegeben:

Datentyp	Bits	$q$	$p$
<b>float</b>	32	8	23
<b>double</b>	64	11	52
<b>long double</b>		$\geq 15$	$\geq 63$

- Festzulegen ist hier nur, ob die Binärrepräsentierung in *little* oder *big endian* übertragen wird.

- Kommunikation ist entweder bilateral (der Normalfall) oder richtet sich an viele Teilnehmer gleichzeitig (*multicast*, *broadcast*).
- Bei einer bilateralen Kommunikation ergibt sich aus der Asymmetrie der Verbindungsaufnahme eine Rollenverteilung, typischerweise die eines Klienten und die eines Diensteanbieters.
- Diese Rollenverteilung bezieht sich immer auf eine konkrete Verbindung, d.h. es können zwischen zwei Kommunikationspartnern mehrere Verbindungen unterschiedlichen Rollenverteilungen bestehen.
- Über ein Protokoll wird geregelt, wie die einzelnen Mitteilungen aussehen und in welcher Abfolge diese gesendet werden dürfen.



- Klassischerweise existieren Netzwerkdienste, die angerufen werden können:
  - ▶ Diese sind sinnvoll, wenn die gleichen Aufgaben wiederkehrend zu lösen sind.
  - ▶ Es bleibt aber das Problem, wie diese Dienste gefunden werden und wie eine sinnvolle Lastverteilung zwischen konkurrierenden Anwendungen erfolgt.
- Bei variierenden Aufgabenstellungen muss ggf. eine Anwendung erst auf genügend Rechnerressourcen verteilt werden:
  - ▶ Wie erfolgt die Verteilung?
  - ▶ Wird die Umfang der Ressourcen zu Beginn oder erst im Laufe der Anwendung festgelegt?
  - ▶ Wie erfolgt die Verbindungsaufnahme untereinander?

- MPI (*Message Passing Interface*) ist ein Standard für eine Bibliotheksschnittstelle für parallele Programme.
- 1994 entstand die erste Fassung (1.0), 1995 die Version 1.2 und seit 1997 gibt es 2.0. Kürzlich (im September 2009) erschien die Version 2.2. Die Standards sind öffentlich unter <http://www.mpi-forum.org/>
- Der Standard umfasst die sprachspezifischen Schnittstellen für Fortran, C und C++.
- Es stehen mehrere Open-Source-Implementierungen zur Verfügung:
  - ▶ OpenMPI: <http://www.open-mpi.org/> (wird von Sun verwendet, auf Theseus)
  - ▶ MPICH2: <http://www.mcs.anl.gov/research/projects/mpich2/>
  - ▶ MVAPICH: <http://mvapich.cse.ohio-state.edu/> (spezialisiert auf Infiniband, auf Pacioli)

- Zu Beginn wird mit  $n$  die Zahl der Prozesse festgelegt.
- Jeder Prozess läuft in seinem eigenen Adressraum und hat innerhalb von MPI eine eigene Nummer (*rank*) im Bereich von 0 bis  $n - 1$ .
- Die Kommunikation mit den anderen Prozessen erfolgt über Nachrichten, die entweder an alle gerichtet werden (*broadcast*) oder individuell versandt werden.
- Die Kommunikation kann sowohl synchron als auch asynchron erfolgen.
- Die Prozesse können in einzelne Gruppen aufgesplittet werden.

mpi-simpson.cpp

```
int main(int argc, char** argv) {
    MPI_Init(&argc, &argv);

    int noprocesses; MPI_Comm_size(MPI_COMM_WORLD, &noprocesses);
    int rank; MPI_Comm_rank(MPI_COMM_WORLD, &rank);

    // process command line arguments
    int n; // number of intervals
    if (rank == 0) {
        cmdname = argv[0];
        if (argc > 2) usage();
        if (argc == 1) {
            n = noprocesses;
        } else {
            istringstream arg(argv[1]);
            if (!(arg >> n) || n <= 0) usage();
        }
    }
    // ...

    MPI_Finalize();

    if (rank == 0) {
        cout << setprecision(14) << sum << endl;
    }
}
```

`mpi-simpson.cpp`

```
MPI_Init(&argc, &argv);  
  
int nofprocesses; MPI_Comm_size(MPI_COMM_WORLD, &nofprocesses);  
int rank; MPI_Comm_rank(MPI_COMM_WORLD, &rank);
```

- Im Normalfall starten alle Prozesse das gleiche Programm und beginnen alle mit *main()*. (Es ist auch möglich, verschiedene Programme über MPI zu koordinieren.)
- Erst nach dem Aufruf von *MPI\_Init()* sind weitere MPI-Operationen zulässig.
- *MPI\_COMM\_WORLD* ist die globale Gesamtgruppe aller Prozesse eines MPI-Laufs.
- Die Funktionen *MPI\_Comm\_size* und *MPI\_Comm\_rank* liefern die Zahl der Prozesse bzw. die eigene Nummer innerhalb der Gruppe (immer ab 0, konsekutiv weiterzählend).

mpi-simpson.cpp

```
// process command line arguments
int n; // number of intervals
if (rank == 0) {
    cmdname = argv[0];
    if (argc > 2) usage();
    if (argc == 1) {
        n = nofprocesses;
    } else {
        istringstream arg(argv[1]);
        if (!(arg >> n) || n <= 0) usage();
    }
}
```

- Der Hauptprozess hat den *rank* 0. Nur dieser sollte verwendet werden, um Kommandozeilenargumente auszuwerten und/oder Ein- und Ausgabe zu betreiben.

mpi-simpson.cpp

```
// broadcast number of intervals  
MPI_Bcast(&n, 1, MPI_INT, 0, MPI_COMM_WORLD);
```

- Mit der Funktion *MPI\_Bcast* kann eine Nachricht an alle Mitglieder einer Gruppe versandt werden.
- Die Funktion bezieht sich auf eine Gruppe, wobei *MPI\_COMM\_WORLD* die globale Gesamtgruppe repräsentiert.
- Der erste Parameter ist ein Zeiger auf das erste zu übermittelnde Objekt. Der zweite Parameter nennt die Zahl der zu übermittelnden Objekte (hier nur 1).
- Der dritte Parameter spezifiziert den Datentyp eines zu übermittelnden Elements. Hier wird *MPI\_INT* verwendet, das dem Datentyp **int** entspricht.
- Der letzte Parameter legt fest, welcher Prozess den Broadcast verschickt. Alle anderen Prozesse, die den Aufruf ausführen, empfangen das Paket.

mpi-simpson.scp

```
// broadcast number of intervals
MPI_Bcast(&n, 1, MPI_INT, 0, MPI_COMM_WORLD);

double value = 0; // summed up value of our intervals;
if (rank < n) {
    int nofintervals = n / nofprocesses;
    int remainder = n % nofprocesses;
    int first_interval = rank * nofintervals;
    if (rank < remainder) {
        ++nofintervals;
        if (rank > 0) first_interval += rank;
    } else {
        first_interval += remainder;
    }
    int next_interval = first_interval + nofintervals;

    double xleft = a + first_interval * (b - a) / n;
    double x = a + next_interval * (b - a) / n;
    value = simpson(f, xleft, x, nofintervals);
}
double sum;
MPI_Reduce(&value, &sum, 1, MPI_DOUBLE, MPI_SUM, 0, MPI_COMM_WORLD);
```



mpi-simpson.cpp

```
double sum;  
MPI_Reduce(&value, &sum, 1, MPI_DOUBLE, MPI_SUM, 0, MPI_COMM_WORLD);
```

- Mit der Funktion *MPI\_Reduce* werden die einzelnen Ergebnisse aller Prozesse (einschließlich dem auswertenden Prozess) eingesammelt und dann mit einer auszuwählenden Funktion aggregiert.
- Der erste Parameter ist ein Zeiger auf ein Einzelresultat. Der zweite Parameter verweist auf die Variable, wo der aggregierte Wert abzulegen ist.
- Der dritte Parameter liegt wieder die Zahl der Elemente fest (hier 1) und der vierte den Datentyp (hier *MPI\_DOUBLE* für **double**).
- Der fünfte Parameter spezifiziert die aggregierende Funktion (hier *MPI\_SUM* zum Aufsummieren) und der sechste Parameter gibt an, welcher Prozess den aggregierten Wert erhält.

MPI unterstützt folgende Datentypen von C++:

<i>MPI_CHAR</i>	<b>signed char</b>
<i>MPI_SIGNED_CHAR</i>	<b>signed char</b>
<i>MPI_UNSIGNED_CHAR</i>	<b>unsigned char</b>
<i>MPI_SHORT</i>	<b>signed short</b>
<i>MPI_INT</i>	<b>signed int</b>
<i>MPI_LONG</i>	<b>signed long</b>
<i>MPI_UNSIGNED_SHORT</i>	<b>unsigned short</b>
<i>MPI_UNSIGNED</i>	<b>unsigned int</b>
<i>MPI_UNSIGNED_LONG</i>	<b>unsigned long</b>
<i>MPI_FLOAT</i>	<b>float</b>
<i>MPI_DOUBLE</i>	<b>double</b>
<i>MPI_LONG_DOUBLE</i>	<b>long double</b>
<i>MPI_WCHAR</i>	<i>wchar_t</i>
<i>MPI_BOOL</i>	<i>bool</i>
<i>MPI_COMPLEX</i>	<i>Complex&lt;float&gt;</i>
<i>MPI_DOUBLE_COMPLEX</i>	<i>Complex&lt;double&gt;</i>
<i>MPI_LONG_DOUBLE_COMPLEX</i>	<i>Complex&lt;long double&gt;</i>

Makefile

```
CXX :=      mpiCC
CXXFLAGS := -fast -library=stlport4
CC :=      mpiCC
CFLAGS :=  -fast -library=stlport4
```

- Die Option *mpi* sollte in *~/.options* genannt werden. Ggf. hinzufügen und erneut anmelden.
- Dann ist */opt/SUNWhpc/HPC8.2.1c/sun/bin* relativ weit vorne im Suchpfad.
- Statt den C++-Compiler von Sun mit *CC* direkt aufzurufen, wird stattdessen *mpiCC* verwendet, das alle MPI-spezifischen Header-Dateien und Bibliotheken automatisch zugänglich macht.
- Die Option *-fast* schaltet alle Optimierungen ein. Die Warnung, die deswegen ausgegeben wird, kann ignoriert werden.

```
theseus$ f
Makefile mpi-simpson.C
theseus$ make mpi-simpson
mpiCC -fast      mpi-simpson.C  -o mpi-simpson
CC: Warning: -xarch=native has been explicitly specified, or implicitly specified
theseus$ time mpirun -np 1 mpi-simpson 10000000
3.1415926535902

real    0m0.95s
user    0m0.87s
sys     0m0.03s
theseus$ time mpirun -np 4 mpi-simpson 10000000
3.1415926535897

real    0m0.39s
user    0m1.02s
sys     0m0.14s
theseus$
```

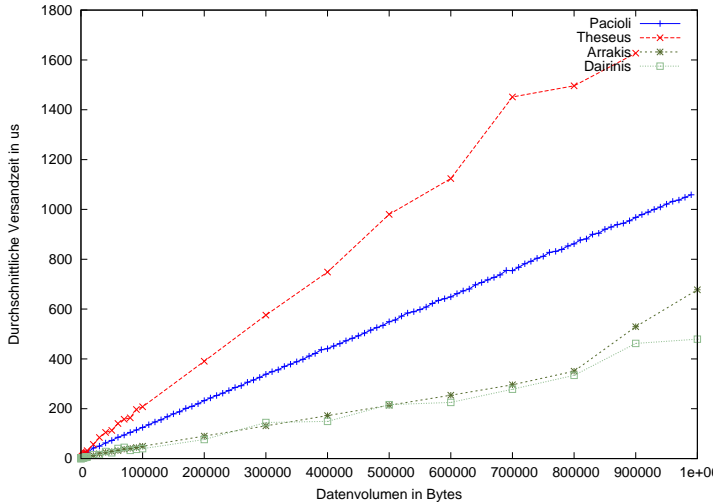
- Mit *mpirun* können MPI-Anwendungen gestartet werden.
- Wenn das Programm ohne *mpirun* läuft, dann gibt es nur einen einzigen Prozess.
- Die Option *-np* spezifiziert die Zahl der zu startenden Prozesse. Per Voreinstellung starten die alle auf der gleichen Maschine.

```
theseus$ cat my-machines
malaga
rom
prag
lille
theseus$ time mpirun -hostfile my-machines -np 4 mpi-simpson 10000000
3.1415926535897

real    0m3.03s
user    0m0.31s
sys     0m0.06s
theseus$
```

- Die Option *-hostfile* ermöglicht die Spezifikation einer Datei mit Rechnernamen. Diese Datei sollte soviel Einträge enthalten wie Prozesse gestartet werden.
- Bei OpenMPI werden die Prozesse auf den anderen Rechnern mit Hilfe der *ssh* gestartet. Letzteres sollte ohne Passwort möglich sein. Entsprechend sollte mit *ssh-keygen* ein Schlüsselpaar erzeugt werden und der eigene öffentliche Schlüssel in *~/.ssh/authorized\_keys* integriert werden.
- Das reguläre Ethernet mit TCP/IP ist jedoch langsam!

- Nach dem Anmelden die Kommandos *module add sge* und *module add mvapich2/gcc/64/1.4.1* ausführen. (mvapich2 ist eine auf das Infini-Band spezialisierte MPI-Implementierung.)
- Danach ist *mpicxx* im Pfad und kann zum Übersetzen verwendet werden. Die sonstigen Optionen entsprechen denen des *g++*. Insbesondere empfiehlt sich *-O3*.
- Weitere Hinweise sind hier zu finden:  
<http://www.mathematik.uni-ulm.de/numerik/teaching/ss11/ScientificComputing/pacioli-mpi.html>



- Pacioli: 8 Prozesse, Infiniband. Gemeinsamer Speicher: Theseus: 6 Prozesse; Arrakis: 2 Prozesse (AMD-Opteron Dual-Core, 3 GHz); Dairinis: 4 Prozesse (Intel Quad-Core, 2,5 GHz)

Warum schneidet die Pacioli mit dem Infiniband besser als die Theseus ab?

- ▶ OpenMPI nutzt zwar gemeinsame Speicherbereiche zur Kommunikation, aber dennoch müssen die Daten beim Transfer zweifach kopiert werden.
- ▶ Das Kopieren erfolgt zu Lasten der normalen CPUs.
- ▶ Hier wäre OpenMP grundsätzlich wesentlich schneller, da dort der doppelte Kopieraufwand entfällt.
- ▶ Sobald kein nennenswerter Kopieraufwand notwendig ist, dann sieht die Theseus mit ihren niedrigeren Latenzzeiten besser aus:  $2,2 \mu s$  vs.  $4,8 \mu s$  bei Pacioli. (Arrakis:  $0.8 \mu s$ ; Dairinis:  $0.9 \mu s$ ).



- Bei inhomogenen Rechnerleistungen oder bei einer inhomogenen Stückelung in Einzelaufgaben kann es sinnvoll sein, die Last dynamisch zu verteilen.
- In diesem Falle übernimmt ein Prozess die Koordination, indem er Einzelaufträge vergibt, die Ergebnisse aufammelt und – sofern noch mehr zu tun ist – weitere Aufträge verschickt.
- Die anderen Prozesse arbeiten alle als Sklaven, die Aufträge entgegennehmen, verarbeiten und das Ergebnis zurücksenden.
- Dies wird an einem Beispiel der Matrix-Vektor-Multiplikation demonstriert, wobei diese Technik in diesem konkreten Fall nicht viel bringt.

```
int main(int argc, char** argv) {
    MPI_Init(&argc, &argv);

    int rank; MPI_Comm_rank(MPI_COMM_WORLD, &rank);
    int nofslaves; MPI_Comm_size(MPI_COMM_WORLD, &nofslaves);
    --nofslaves; assert(nofslaves > 0);

    if (rank == 0) {
        int n; double** A; double* x;
        if (!read_parameters(n, A, x)) {
            cerr << "Invalid input!" << endl;
            MPI_Abort(MPI_COMM_WORLD, 1);
        }
        double* y = new double[n];
        gemv_master(n, A, x, y, nofslaves);
        for (int i = 0; i < n; ++i) {
            cout << " " << y[i] << endl;
        }
    } else {
        gemv_slave();
    }

    MPI_Finalize();
}
```

```
static void gemv_slave() {
    int n;
    MPI_Bcast(&n, 1, MPI_INT, 0, MPI_COMM_WORLD);
    double* x = new double[n];
    MPI_Bcast(x, n, MPI_DOUBLE, 0, MPI_COMM_WORLD);

    double* row = new double[n];
    // receive tasks and process them
    for(;;) {
        // receive next task
        MPI_Status status;
        MPI_Recv(row, n, MPI_DOUBLE, 0, MPI_ANY_TAG,
                 MPI_COMM_WORLD, &status);
        if (status.MPI_TAG == FINISH) break;
        // process it
        double result = 0;
        for (int i = 0; i < n; ++i) {
            result += row[i] * x[i];
        }
        // send result back to master
        MPI_Send(&result, 1, MPI_DOUBLE, 0, 0, MPI_COMM_WORLD);
    }
    // release allocated memory
    delete[] x; delete[] row;
}
```

mpi-gemv.cpp

```
int n;  
MPI_Bcast(&n, 1, MPI_INT, 0, MPI_COMM_WORLD);  
double* x = new double[n];  
MPI_Bcast(x, n, MPI_DOUBLE, 0, MPI_COMM_WORLD);
```

- Zu Beginn werden die Größe des Vektors und der Vektor selbst übermittelt.
- Da alle Sklaven den gleichen Vektor (mit unterschiedlichen Zeilen der Matrix) multiplizieren, kann der Vektor ebenfalls gleich zu Beginn mit *Bcast* an alle verteilt werden.

mpi-gemv.cpp

```
MPI_Status status;  
MPI_Recv(row, n, MPI_DOUBLE, 0, MPI_ANY_TAG,  
         MPI_COMM_WORLD, &status);  
if (status.MPI_TAG == FINISH) break;
```

- Mit *MPI\_Recv* wird hier aus der globalen Gruppe eine Nachricht empfangen.
- Die Parameter: Zeiger auf den Datenpuffer, die Zahl der Elemente, der Element-Datentyp, der sendende Prozess, die gewünschte Art der Nachricht (*MPI\_ANY\_TAG* akzeptiert alles), die Gruppe und der Status, über den Nachrichtenart ermittelt werden kann.
- Nachrichtenarten gibt es hier zwei: *NEXT\_ROW* für den nächsten Auftrag und *FINISH*, wenn es keine weiteren Aufträge mehr gibt.

mpi-gemv.cpp

```
MPI_Send(&result, 1, MPI_DOUBLE, 0, 0, MPI_COMM_WORLD);
```

- *MPI\_Send* versendet eine individuelle Nachricht synchron, d.h. diese Methode kehrt erst dann zurück, wenn der Empfänger die Nachricht erhalten hat.
- Die Parameter: Zeiger auf den Datenpuffer, die Zahl der Elemente (hier 1), der Element-Datentyp, der Empfänger-Prozess (hier 0) und die Art der Nachricht (0, spielt hier keine Rolle).

```
static void
gemv_master(int n, double** A, double *x, double* y,
            int nofslaves) {
    // broadcast parameters that are required by all slaves
    MPI_Bcast(&n, 1, MPI_INT, 0, MPI_COMM_WORLD);
    MPI_Bcast(x, n, MPI_DOUBLE, 0, MPI_COMM_WORLD);

    // send out initial tasks for all slaves
    int* tasks = new int[nofslaves];
    // ...

    // collect results and send out remaining tasks
    // ...

    // release allocated memory
    delete[] tasks;
}
```

- Zu Beginn werden die beiden Parameter  $n$  und  $x$ , die für alle Sklaven gleich sind, mit *Bcast* verteilt.
- Danach erhält jeder der Sklaven einen ersten Auftrag.
- Anschließend werden Ergebnisse eingesammelt und – sofern noch etwas zu tun übrig bleibt – die Anschlußaufträge verteilt.

mpi-gemv.cpp

```
// send out initial tasks for all slaves
// remember the task for each of the slaves
int* tasks = new int[nofslaves];
int next_task = 0;
for (int slave = 1; slave <= nofslaves; ++slave) {
    if (next_task < n) {
        int row = next_task++; // pick next remaining task
        MPI_Send(A[row], n, MPI_DOUBLE, slave, NEXT_ROW,
                 MPI_COMM_WORLD);
        // remember which task was sent out to whom
        tasks[slave-1] = row;
    } else {
        // there is no work left for this slave
        MPI_Send(0, 0, MPI_DOUBLE, slave, FINISH, MPI_COMM_WORLD);
    }
}
```

- Die Sklaven erhalten zu Beginn jeweils eine Zeile der Matrix  $A$ , die sie dann mit  $x$  multiplizieren können.



```
// collect results and send out remaining tasks
int done = 0;
while (done < n) {
    // receive result of a completed task
    double value = 0; // initialize it to get rid of warning
    MPI_Status status;
    MPI_Recv(&value, 1, MPI_DOUBLE,
             MPI_ANY_SOURCE, MPI_ANY_TAG, MPI_COMM_WORLD, &status);
    int slave = status.MPI_SOURCE;
    int row = tasks[slave-1];
    y[row] = value;
    ++done;
    // send out next task, if there is one left
    if (next_task < n) {
        row = next_task++;
        MPI_Send(A[row], n, MPI_DOUBLE, slave, NEXT_ROW,
                 MPI_COMM_WORLD);
        tasks[slave-1] = row;
    } else {
        // send notification that there is no more work to be done
        MPI_Send(0, 0, MPI_DOUBLE, slave, FINISH, MPI_COMM_WORLD);
    }
}
```

`mpi-gemv.cpp`

```
MPI_Status status;  
MPI_Recv(&value, 1, MPI_DOUBLE,  
        MPI_ANY_SOURCE, MPI_ANY_TAG, MPI_COMM_WORLD, &status);  
int slave = status.MPI_SOURCE;
```

- Mit *MPI\_ANY\_SOURCE* wird angegeben, dass ein beliebiger Sender akzeptiert wird.
- Hier ist die Identifikation des Sklaven wichtig, damit das Ergebnis korrekt in *y* eingetragen werden kann. Dies erfolgt hier mit *status.Get\_source()*.

```
int MPI_Send(void* buf, int count,  
             MPI_Datatype datatype,  
             int dest, int tag, MPI_Comm comm);
```

- MPI-Nachrichten bestehen aus einem Header und der zu versendenden Datenstruktur (*buf*, *count* und *datatype*).
- Der (sichtbare) Header ist ein Tupel bestehend aus der
  - ▶ Kommunikationsdomäne (normalerweise *MPI\_COMM\_WORLD*), dem
  - ▶ Absender (*rank* innerhalb der Kommunikationsdomäne) und einer
  - ▶ Markierung (*tag*).

```
int MPI_Recv(void* buf, int count,  
             MPI_Datatype datatype,  
             int source, int tag, MPI_Comm comm,  
             MPI_Status* status);
```

Eine mit *MPI\_Send* versendete MPI-Nachricht passt zu einem *MPI\_Recv* beim Empfänger, falls gilt:

- ▶ die Kommunikationsdomänen stimmen überein,
- ▶ der Absender stimmt mit *source* überein oder es wurde *MPI\_ANY\_SOURCE* angegeben,
- ▶ die Markierung stimmt mit *tag* überein oder es wurde *MPI\_ANY\_TAG* angegeben,
- ▶ die Datentypen sind identisch und
- ▶ die Zahl der Elemente ist kleiner oder gleich der angegebenen Buffergröße.

- Wenn die Gegenseite bei einem passenden *MPI\_Recv* auf ein Paket wartet, werden die Daten direkt übertragen.
- Wenn die Gegenseite noch nicht in einem passenden *MPI\_Recv* wartet, **kann** die Nachricht gepuffert werden. In diesem Falle wird „im Hintergrund“ darauf gewartet, dass die Gegenseite eine passende *MPI\_Recv*-Operation ausführt.
- Alternativ kann *MPI\_Send* solange blockieren, bis die Gegenseite einen passenden *MPI\_Recv*-Aufruf absetzt.
- Wird die Nachricht übertragen oder kommt es zu einer Pufferung, so kehrt *MPI\_Send* zurück. D.h. nach dem Aufruf von *MPI\_Send* kann in jedem Falle der übergebene Puffer andersweitig verwendet werden.
- Die Pufferung ist durch den Kopieraufwand teuer, ermöglicht aber die frühere Fortsetzung des sendenden Prozesses.
- Ob eine Pufferung zur Verfügung steht oder nicht und welche Kapazität sie ggf. besitzt, ist systemabhängig.

mpi-deadlock.cpp

```
int main(int argc, char** argv) {
    MPI_Init(&argc, &argv);
    int noprocesses; MPI_Comm_size(MPI_COMM_WORLD, &noprocesses);
    int rank; MPI_Comm_rank(MPI_COMM_WORLD, &rank);
    assert(noprocesses == 2); const int other = 1 - rank;
    const unsigned int maxsize = 8192;
    double* bigbuf = new double[maxsize];
    for (int len = 1; len <= maxsize; len *= 2) {
        MPI_Send(bigbuf, len, MPI_DOUBLE, other, 0, MPI_COMM_WORLD);
        MPI_Status status;
        MPI_Recv(bigbuf, len, MPI_DOUBLE, other, 0, MPI_COMM_WORLD,
            &status);
        if (rank == 0) cout << "len = " << len << " survived" << endl;
    }
    MPI_Finalize();
}
```

- Hier versuchen die beiden Prozesse 0 und 1 sich erst jeweils etwas zuzusenden, bevor sie *MPI\_Recv* aufrufen. Das kann nur mit Pufferung gelingen.

```
dairinis$ mpirun -np 2 mpi-deadlock
len = 1 survived
len = 2 survived
len = 4 survived
len = 8 survived
len = 16 survived
len = 32 survived
len = 64 survived
len = 128 survived
len = 256 survived
^Cmpirun: killing job...

-----
mpirun noticed that process rank 0 with PID 28203 on node dairinis exited on signal 0 (UNKNOWN SIGNAL).
-----
2 total processes killed (some possibly by mpirun during cleanup)
mpirun: clean termination accomplished

dairinis$
```

- Hier war die Pufferung nicht in der Lage, eine Nachricht mit 512 Werten des Typs **double** aufzunehmen.
- MPI-Anwendungen, die sich auf eine vorhandene Pufferung verlassen, sind unzulässig bzw. deadlock-gefährdet in Abhängigkeit der lokalen Rahmenbedingungen.