

В.Н. Попов¹, И.А. Ботыгин², В.А. Тартаковский¹

Технология вычислительных кластеров в задачах дендроэкологической диагностики

¹ *Институт мониторинга климатических и экологических систем СО РАН, г. Томск*

² *Томский политехнический университет*

Поступила в редакцию 12.11.2004 г.

Представлен многомашинный программный комплекс, реализующий распределенное решение задач дендроэкологической диагностики. В архитектуре комплекса выделены вычислительный кластер и коммуникационный сервер, обеспечивающие доступ к информационным материалам в области дендроэкологии и инструментальным средствам для анализа годичных колец деревьев с целью выявления изменений параметров окружающей среды, отраженных в приросте дерева. В основу моделирования годичных колец деревьев положены фундаментальные свойства радиального роста дерева – монотонность роста во времени и пространстве и ограниченность его скорости, а изменения плотности древесины рассматриваются как пространственно-временной колебательный процесс.

Введение

В настоящее время новые информационные технологии активно используются и в задачах диагностики состояния окружающей среды и экологии человека. В частности, возрастают роль долгосрочного экологического прогнозирования и значение дендроэкологических исследований, реализованных и доступных в компьютерных сетях. Фундаментальные академические исследования в области дендроэкологии требуют разработки новых методов получения экспериментальных данных, которые бы более точно отражали изменение соответствующих климатических данных.

Особенность задач дендроэкологической диагностики заключается в необходимости математической обработки очень большого объема данных (иногда требуется обработка временных рядов наблюдений за 1000 лет и более) и коллективной (одновременной) работы многих сотрудников практически на всех этапах дендроэкологических исследований, а также в необходимости хранения и систематизации больших объемов неоднородной структурированной информации (собственно хронологические ряды наблюдений, результаты обработки, сопутствующие метеорологические, геологические, геофизические, аэрокосмические и т.п. ряды наблюдений). Таким образом, большой объем вычислений, обычный при обработке дендроэкологических данных, подразумевает использование высокопроизводительных вычислительных систем и разработку новых распределенных методов их обработки. Методы распределенной обработки подразумевают физическое распределение обработки информации в пространстве на нескольких вычислительных машинах, которые связаны между собой каналами передачи данных.

Технологические принципы, положенные в основу распределенной обработки, – это, по сути, принципы реализации открытых систем [1, 2], такие как производительность (Performance), масштабируемость (Scalability), мобильность программного обеспечения (Software portability), надежность программных средств (Software reliability), защита данных (Data security), защита ЭВМ (Computer security). Именно реализация этих принципов обеспечивает необходимые для практического применения и коллективную работу пользователей, и мобильность (переносимость) программного обеспечения, и совместное функционирование с уже разработанными программными средствами, и защиту конфиденциальной информации, и многое другое. На решение задач обработки большого объема данных направлены исследования в области кластеризации вычислительных компонентов (вычислительные кластеры) [3] и grid-технологии [4, 5]. Исследования в области виртуализации центров обработки данных [6] направлены на обеспечение коллективного (множественного) доступа к вычислительным ресурсам. А концепция хранилищ данных направлена на решение проблемы хранения и использования неоднородной информации [7, 8].

В данной статье описывается многомашинный комплекс обработки дендроэкологических данных, направленный на распределенную обработку и анализ изображений годичных колец деревьев с целью выявления изменений параметров окружающей среды, отраженных в приросте дерева. Традиционные дендрохронологические ряды ориентированы на датировку событий и содержат информацию только о ширине годичного кольца. Известные методы измерения роста деревьев трудоемки и не имеют достаточной алгоритмической и программной поддержки. В пределах вегетационного периода

с одного дерева можно отобрать только небольшое число дискретных проб, поэтому качество полученной реализации определяется только по завершению роста. Также существуют проблемы формирования однородной выборки с достаточным объемом. Для изучения статистических взаимосвязей процесса роста и погоды необходима фиксация этих случайных процессов как функций времени в период вегетации [9, 10]. Предложено рассматривать изменение плотности древесины вдоль радиуса диска дерева как некоторое колебание, фаза которого является строго возрастающей функцией радиуса. Радиальный рост определен как монотонная функция времени, обратная по отношению к фазе. При этом влияние внешней среды и генетическая программа дерева проявляются в форме амплитудно-фазовой модуляции клеточной структуры. Используемый математический аппарат основан на введенном условии дисперсионной причинности и формализме аналитического сигнала [11].

1. Архитектура вычислительного кластера

Отмеченные выше особенности задач дендрологии создают определенные проблемы при формировании массовых профессиональных центров обработки и хранения дендрологических данных в отдельно взятом научном учреждении. Слишком большие вычислительные возможности и мощные базы данных требуются для этого, и очень

велики материальные затраты. Естественным решением в сложившейся ситуации является концентрация информационного, программного и технического обеспечения, решение задач дендрологии в нескольких специализированных суперкомпьютерных центрах и виртуализация доступа исследователей к этим информационно-обрабатывающим ресурсам.

Основопологающей концепцией реализации многомашинного комплекса обработки дендрологических данных является его функционирование в сетевом варианте. Учитывая международное значение подобных центров обработки, а также слабую инфраструктуру глобальных телекоммуникаций в России, в качестве основы для реализации сетевой работы комплекса выбран web-интерфейс.

Авторами были проведены экспериментальные исследования требуемой вычислительной мощности на такой задаче в области дендрологии, как обработка изображений годовичных колец деревьев (сканирование, обработка фильтрами). В зависимости от размера обрабатываемого изображения требовалось до нескольких десятков минут процессорного времени (Celeron 1,7 МГц, DDR RAM 256 Мбайт). Нагрузка на процессор возрастала практически пропорционально числу подключившихся пользователей. Одним из решений проблемы недостаточности вычислительных ресурсов является объединение нескольких компьютеров в многомашинный комплекс обработки (рис. 1), так как нереально увеличивать вычислительную мощность одного компьютера до бесконечности.

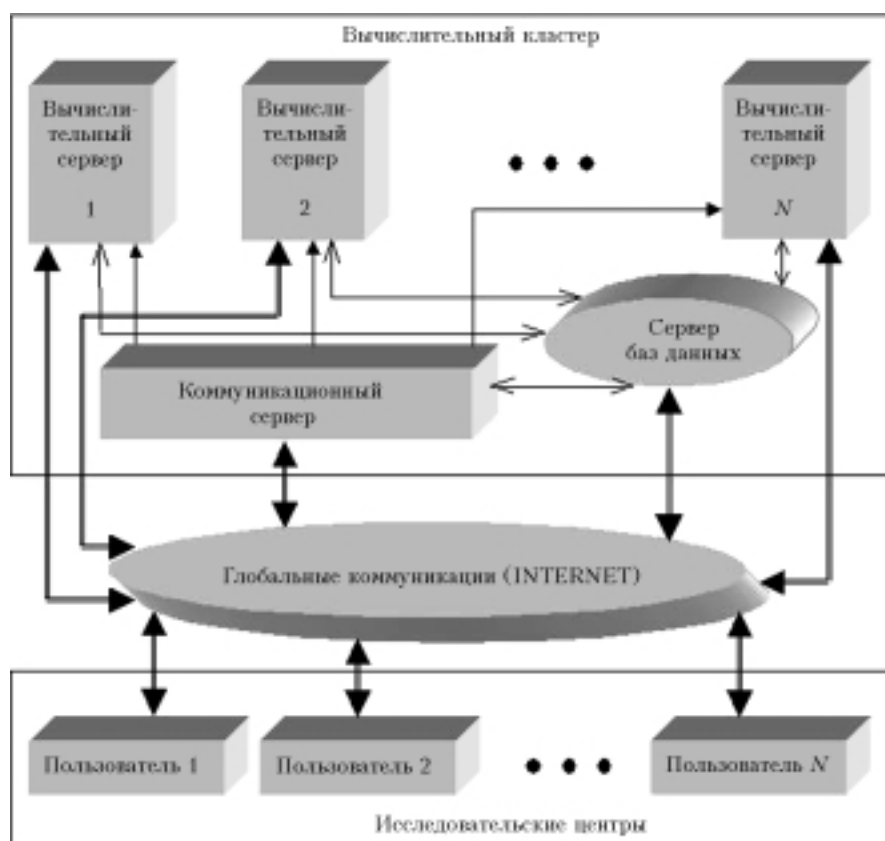


Рис. 1. Архитектура многомашинного комплекса обработки дендрологических данных

Для обеспечения взаимодействия серверов (коммуникационный сервер ↔ вычислительный сервер, коммуникационный сервер ↔ сервер базы данных, вычислительный сервер ↔ сервер базы данных) внутри вычислительного кластера, в случае использования на каком-либо сервере сетевого фильтра, необходимо включить IP-адреса соответствующих серверов в список разрешенных входящих и исходящих TCP-соединений этого сетевого фильтра.

Планирование и диспетчирование процессов обработки дендрологических данных возложены на специальный коммуникационный сервер. Взаимодействие всех пользователей с комплексом обработки осуществляется только через коммуникационный сервер. Непосредственно на этом сервере математические вычисления не выполняются. Основной задачей коммуникационного сервера является обеспечение оптимальной загрузки имеющихся в его распоряжении вычислительных серверов и режима работы в реальном времени (on line) пользователей.

Задачи, на решение которых направлены алгоритмы математической обработки, реализованные на вычислительных серверах кластера, могут быть самыми различными. Например, вычисление радиального роста дерева и плотности древесины в период вегетации, построение длительных древесно-кольцевых хронологий и т.д. Важно, чтобы эти системы обработки дендрологических данных придерживались единого интерфейса для связи с коммуникационным сервером.

Для полноценного функционирования комплекса обработки дендрологических данных необходимо организовать автоматизированный сбор, систематизацию и хранение научной информации в области дендрологического мониторинга, а также формирование и ведение базы дендрологических данных. Ни одну из этих задач невозможно решить без применения системы управления базами данных (СУБД). Для универсальности программного обеспечения комплекса необходимо, чтобы СУБД понимала язык запросов SQL. Хотя и есть определенные нюансы в использовании команд SQL различных баз данных, но часто они сводятся к различию в синтаксисе некоторых команд и их опций. Ключевые же команды SELECT, INSERT, UPDATE и DELETE, которые служат для выбора, добавления, обновления и удаления информации, обычно стандартны. В принципе, необходимо, чтобы комплекс обеспечивал взаимодействие с любыми серверами баз данных, понимающих язык запросов SQL. Типы данных, которые необходимо хранить в реляциях (отношениях) СУБД, должны включать целочисленные и вещественные значения различной размерности, фиксированные и динамические строки, даты, Blob-поля для хранения произвольных данных, например бинарных файлов изображений.

Как и многие прикладные многопользовательские системы, обработка дендрологических данных все настойчивее требует перехода к модели вычислений «клиент — сервер» и распределенной обработке. Применительно к комплексу анализа дендрологических данных архитектура «клиент —

сервер» интересна и актуальна, главным образом, потому, что обеспечивает простые системные программные решения и экономически относительно дешевое решение проблемы коллективного удаленного доступа пользователей как к вычислительным, так и к информационным ресурсам комплекса. Отметим, что наиболее оптимальная реализация клиентской части основывается на использовании стандартных средств навигации в сети Internet (Internet Explorer, Netscape Navigator и других подобных браузеров), а не на разработке индивидуальных клиентских частей.

2. Математическая модель годичных колец деревьев

Математическую модель годичных колец можно представить в виде интерферограммы, образованной «экологическим» $U(\rho, \theta)$ и «биологическим» $B(\rho, \theta)$ полями:

$$R(\rho, \theta) = |B(\rho, \theta) + U(\rho, \theta)|^2 = u^2 + b^2 + 2ub \cos \Phi, \quad (1)$$

где ρ, θ — полярные координаты; $b = |B|$; $u = |U|$; $\Phi = \arg B - \arg U$. Эта разность $\Phi(\rho, \theta)$ содержит информацию о параметрах окружающей среды.

Фундаментальным свойством функции $\Phi(\rho, \theta)$ является монотонность в интервале $[0, \rho]$, что обеспечивает наличие концентрических кольцеобразных структур вокруг центра $\rho = 0$. При этом также предполагается медленное изменение модулей $b(\rho, \theta)$ и $u(\rho, \theta)$ на среднем периоде кольцевой структуры. Более того, из этих свойств следует, что интерференционный компонент [третье слагаемое в (1)] является действительной частью аналитического сигнала по переменной ρ .

Традиционный метод демодуляции кольцевых структур (восстановление функции $\Phi(\rho, \theta)$) состоит в прослеживании колец, определении их порядка или нумерации. Но существуют и другие алгоритмы, не требующие явного анализа структуры колец. При наличии аналитического сигнала функция $\Phi(\rho, \theta)$ восстанавливается с использованием преобразования Гильберта \mathbf{H} и полосовой фильтрации \mathbf{F} по переменной ρ для всех θ [12]:

$$\Phi(\rho, \theta) = \arg[\mathbf{F} R(\rho, \theta)] = \arctg\{[\mathbf{HF} R(\rho, \theta)] / \mathbf{F} R(\rho, \theta)\}. \quad (2)$$

Смысл функции $\Phi(\rho, \theta)$ в дендрологии определяется ее составляющими:

$$\Phi(\rho, \theta) = 2\pi\rho^n/\tau + g_b(\rho) + g_u(\rho) + s(\rho, \theta), \quad (3)$$

где первое слагаемое определяет степенной рост дерева со средним периодом кольца τ в отсутствие каких-либо возмущений, второе — g_b — возрастной тренд и другие изотропные эндогенные факторы, третье — g_u есть изотропные экзогенные факторы

и, наконец, функция $s(\rho, \theta)$ определяет влияние солнца, наклона земной оси, широты места и розы ветров. Для разделения этих составляющих и определения их характеристик в присутствии шумов применяются фильтрация: медианная, полиномиальная и в тригонометрическом базисе, а также вычисление интегральных моментов функции $R(\rho, \theta)$.

3. Программная реализация многомашинного комплекса обработки дендроэкологических данных

Рассмотренная выше схема кластеризации вычислительных ресурсов была реализована для решения задачи анализа изображений годичных колец деревьев. Для ускорения процесса вычислений все необходимые математические преобразования были реализованы в среде серверного приложения. Технология OLE (Object Linking and Embedding) для распределения обработки по компонентам сети (связывания и выполнения программ в других вычислительных средах) не использовалась.

На вычислительный кластер возложено выполнение следующих задач.

Сбор, систематизация и хранение научной информации в области дендроэкологического мониторинга. Информация включает как ссылки на соответствующие материалы, в том числе и web-узлы, где можно найти научную информацию, так и полнотекстовые документы. Также предусмотрены возможность удаленной пересылки файлов непосредственно на сервер и внесение их в базу данных.

Формирование и ведение базы экспериментальных дендроэкологических данных. База создается посредством проведения на сервере модельных экспериментов на основе обработки изображений спилов деревьев (графических файлов). Вместе с выходной информацией, полученной в результате модельного эксперимента, в базу данных заносится информация об обработанном диске дерева: порода дерева, его ориентация, место сбора, широта и долгота места, климатические данные. Непосредственная работа с базой данных предусмотрена через специально разработанный для этих целей интерфейс сервера. Также возможно автономное формирование этой базы через интерфейс, предоставленный администратору.

Анализ данных в области дендроэкологии. Из спектра всех задач рассматривается задача анализа годичных колец деревьев, в частности: вычисление азимута и среднеквадратической ширины области максимального прироста годичных колец деревьев и вычисление значений индексов прироста ширины годичных колец деревьев. С информационной точки зрения годичные кольца на поперечном спиле дерева представляют собой модулированный сигнал, зависящий от двух переменных (азимут и расстояние от биологического центра диска дерева). Отметим, что именно двумерный анализ годичных колец деревьев дает возможность получить новые

результаты об изотропности климатоэкологического состояния окружающей среды.

Оценка экологического состояния окружающей среды. Оценка основана на информации, содержащейся в годичных слоях древесины, расшифровка которой позволяет выявить сезонные, годичные и многолетние закономерности изменения климата за интервалы времени, намного превышающие длительность метеорологических наблюдений. При этом можно производить реконструкцию многих важных климатических характеристик, основными из которых являются температура воздуха в различные сезоны и за год, количество осадков в различные сезоны и за год, аномалии атмосферного давления, повторяемость и интенсивность засух, повторяемость заморозков в течение вегетационного периода и сильных морозов в зимнее время [13].

В структуре вычислительного кластера выделен специальный коммуникационный сервер. Основной функцией администрирования коммуникационного сервера является формирование базы вычислительных серверов и решаемых на них задач. Информация о вычислительных серверах хранится в специальной таблице базы данных (таблица векторов связи). Полями таблицы являются: универсальный локатор вычислительного сервера (DNS-имя или IP-адрес), идентификатор выполняемых вычислительным сервером задач, количество пользователей, использующих данный вычислительный сервер, максимальное количество одновременно работающих пользователей, состояние вычислительного сервера (ON/OFF или наличие связи).

Для информационной поддержки функционирования многомашинного комплекса обработки дендроэкологических данных использована СУБД Oracle. С сервером баз данных Oracle взаимодействуют как вычислительные серверы, так и коммуникационный сервер. Синхронизация и корректная обработка всех обращений к базе данных осуществляются стандартными (штатными) средствами СУБД Oracle. Серверные части вычислительного кластера, реализующего основную математическую обработку, и коммуникационного сервера выполнены на базе Apache HTTP Server, а клиентская часть, реализующая интерфейс удаленного взаимодействия с пользователем, – MS Internet Explorer.

4. Анализ годичных колец деревьев

В эксперименте по функционированию многомашинного комплекса обработки дендроэкологических данных были задействованы три вычислительных сервера. На них были развернуты описанные выше и реально действующие серверные приложения анализа годичных колец деревьев. На рис. 2 приведены схема взаимодействия и потоки данных пользователя с конкретным вычислительным сервером.

Выбор пользователем класса решаемых задач осуществляется путем настройки на соответствующий пункт меню при установлении связи с коммуникационным сервером.



Рис. 2. Схема взаимодействия и потоки данных пользователя с конкретным вычислительным сервером

Алгоритм выбора вычислительного сервера коммуникационным сервером заключается в следующем: 1) на основании идентификатора указанного пользователем класса задач из таблицы векторов связи формируется список вычислительных серверов, которые могут выполнить эту задачу; 2) из сформированного списка вычислительных серверов вычеркиваются выключенные серверы и серверы, с которыми связь по техническим причинам невозможна; 3) из оставшегося списка вычислительных серверов вычеркиваются серверы, для которых количество пользователей, использующих данный вычислительный сервер, равно максимальному количеству одновременно работающих пользователей; 4) рабочим вычислительным сервером становится тот, у которого разница между количеством пользователей, использующих данный вычислительный сервер, и максимальным количеством одновременно работающих пользователей максимальна. При совпадении этих показателей выбирается первый из списка вычислительный сервер.

После определения вычислительного сервера осуществляется передача управления на этот сервер и инкремент поля количества пользователей, использующих данный вычислительный сервер, в таблице векторов связи.

Последовательность действий пользователя включала:

Вычисление азимута и среднеквадратичной ширины области максимального прироста годовых колец деревьев: 1) определение центра колец;

2) сканирование изображения по всем азимутам; 3) преобразование изображения из полярной системы координат в декартову [14]; 4) обработка изображения медианным (рис. 3,а) и полиномиальным [15] (рис. 3,б) фильтрами; 5) преобразование изображения из градаций серого в черно-белое; 6) расчет ширины годовых колец и интегрального прироста по всем азимутам; 7) нормирование ширины каждого кольца; 8) определение азимута максимального прироста и среднеквадратичной ширины прироста.

На рис. 3,а показано воздействие медианного фильтра с пятиэлементным окном на радиальное сечение диска дерева. На рис. 3,б показано воздействие полиномиального фильтра (полином пятой степени) на радиальное сечение диска дерева.



Рис. 3. Медианная (а) и полиномиальная (б) фильтрации: — — до фильтрации; — — после фильтрации

Вычисление значений индексов прироста ширины годовых колец деревьев: 1) сканирование изображения по выбранному сечению; 2) обработка изображения полиномиальным фильтром; 3) преобразование изображения из градаций серого в черно-белое; 4) вычисление значений индексов прироста по ширине годовых колец деревьев; 5) вычисление среднего значения по индексам прироста годовых колец деревьев [16].

На рис. 4 приведена иллюстрация вычисления значений индексов прироста по данным исследования годовых колец деревьев за 50 лет.

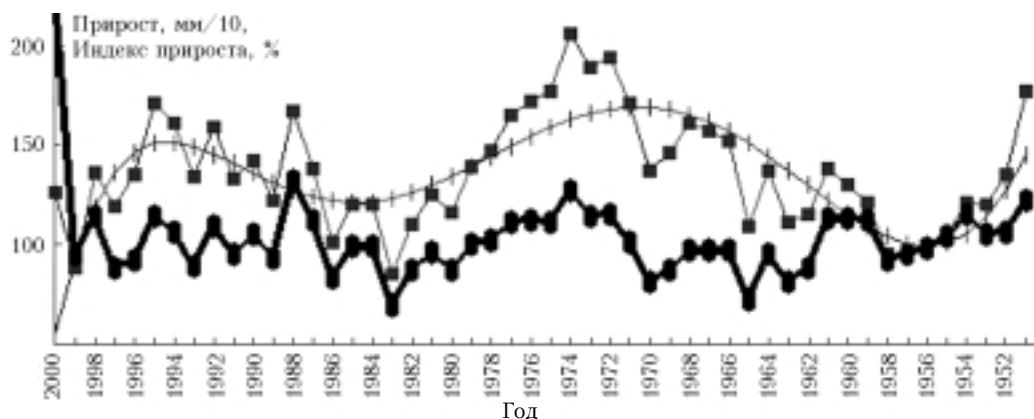


Рис. 4. Вычисление значений индексов прироста годовых колец деревьев: квадраты — прирост; кружочки — индексы прироста; вертикальные черточки — полином

Вычисление значений индексов прироста ширины годичных колец деревьев осуществляется в целях стандартизации представления данных. Стандартизация устраняет из исходных измерений ширины годичных колец возрастную тенденцию, переводя последовательность в безразмерное значение индексов ширины годичных колец. В изменчивости индексов влияние внешних условий максимально и процедуры дальнейшего анализа позволяют выделить ведущие факторы среды и оценить их относительный вклад. Здесь индекс прироста (индекс ширины годичных колец) — трансформированное значение ширины годичных колец, полученное путем деления ее на соответствующее ей значение на сглаженной кривой.

После завершения работы пользователем или его бездействия в течение 10 мин происходят отключение пользователя из системы и декремент поля количества пользователей, использующих данный вычислительный сервер, в таблице векторов связи.

Актуальность и достоверность поля состояния вычислительного сервера (ON/OFF или наличие связи) в таблице векторов связи подтверждались специальной программой, работавшей на коммуникационном сервере с периодичностью 30 мин.

Заключение

Представлена архитектура многомашинного комплекса, включающая вычислительный кластер и коммуникационный сервер, обеспечивающие доступ как к информационным материалам в области дендрэкологии, так и к инструментальным средствам для математического анализа годичных колец деревьев с целью выявления изменений параметров окружающей среды, отраженных в приросте дерева. Математическая обработка дендрэкологических данных (годичных колец деревьев) основана на представлении плотности древесины как пространственно-временном колебательном процессе и включает вычисление азимута и среднеквадратичной ширины области максимального прироста годичных колец деревьев, а также значений индексов прироста ширины годичных колец деревьев.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект № 05-07-98009.

1. Гуляев Ю.В., Олейников А.Я. Открытые системы: от принципов к технологии // Информационные технологии и вычислительные системы. 2003. № 3. С. 4–12.
2. Батоврин В.К., Гуляев Ю.В., Ижванов Ю.Л., Олейников А.Я., Сигов А.С., Тихонов А.Н. Применение технологии открытых систем для построения и исполь-

зования инфраструктуры высокопроизводительных ресурсов для науки и образования. Обращение к документу 02.07.2004. http://www.cplire.ru/win/casr/os/3_12/8/sp2.htm

3. Митрофанов В., Слуцкий А., Ларионов К., Эйсымонт Л. Направления развития отечественных высокопроизводительных систем // Открытые системы. 2003. № 5. Обращение к документу 02.07.2004. <http://www.osp.ru/os/2003/05/029.htm>
4. Лизун А. Grid-технологии в действии // ComputerWorld. 2003. № 14 (407). Обращение к документу 03.07.2004. <http://www.comizdat.com/3/17/19/2524/>
5. Каменчиков М.А., Корниенко В.Н. Grid и технология открытых систем // Информационные технологии и вычислительные системы. 2003. № 3. С. 45–50.
6. Черняк Л. Виртуализация центра обработки данных // Открытые системы. 2004. № 5. Обращение к документу 02.07.2004. <http://www.osp.ru/os/2004/05/031.htm>
7. Сахаров А.А. Концепции построения и реализации информационных систем, ориентированных на анализ данных // Системы управления базами данных. 1996. № 4. С. 55–70.
8. Львов В. Создание систем поддержки принятия решений на основе хранилищ данных // Системы управления базами данных. 1997. № 3. Обращение к документу 03.07.2004. <http://www.isuct.ru/~ivt/books/DBMS/DBMS7/dbms/1997/03/30.htm>
9. Ваганов Е.А., Шашкин А.В. Рост и структура годичных колец хвойных. Новосибирск: Наука, 2000. 232 с.
10. Шиятов С.Г., Ваганов Е.А., Кирдянов А.В., Крулов В.Б., Мазепа В.С., Наурузбаев М.М., Хантемиров Р.М. Методы дендрохронологии. Часть I. Основы дендрохронологии. Сбор и получение древесно-кольцевой информации. Красноярск: Крас.ГУ, 2000. 80 с.
11. Тартаковский В.А., Исаев Ю.Н., Несветайло В.Д., Волков Ю.В., Попов В.Н. Математическая модель радиального сечения годичных колец деревьев // Автоматика. 2003. Т. 38. № 5. С. 118–127.
12. Витриченко В.А., Лукин В.П., Пушиной Л.А., Тартаковский В.А. Проблемы оптического контроля. Новосибирск: Наука, 1990. 351 с.
13. Шиятов С.Г. Дендрэкология, ее принципы и методы // Зап. Свердлов. отдел. Всесоюз. ботан. об-ва. 1973. № 6. С. 53–81.
14. Вельтмандер П.В. Машинная графика. Обращение к документу: 14.02.2002. <http://algotlist.by.ru/graphics/index.html>
15. Форсайт Дж., Малькольм М., Моулдер К. Машинные методы математических вычислений: Пер. с англ. М.: Мир, 1980. 243 с.
16. Попов В.Н., Ботыгин И.А., Волков Ю.В., Тартаковский В.А. Информационная технология анализа годичных колец деревьев: Матер. симпозиум // IV Междунар. симпозиум. «Контроль и реабилитация окружающей среды». Томск, 2002. С. 103–104.

V.N. Popov, I.A. Botygin, V.A. Tartakovsky. **Computing cluster technology in dendroecology diagnostic problems.**

A multicomputer system is introduced. It realizes the distributed solution of dendroecology diagnostics problems. The system architecture includes a computing cluster and the communication server that provide access to information on dendroecology and tools for tree-ring processing for revealing the changes in environmental parameters reflected in tree growth. Fundamental properties of tree radial growth (monotonic growth in time and space and its limited speed) are put in a basis of tree-ring modelling, and changes of wood density are considered as a space-time oscillatory process.