Кластеры Росгидромета: фотопрогулка по ВЦ

Создание прогноза погоды - один из наиболее трудоемких вычислительных процессов, так как за короткое время требуется обработать огромный массив данных. Справится с этим могут только суперкомпьютеры. В 2009 году ГВЦ Росгидромета совместно с Новосибирским и Хабаровским ЦГМС-РСМЦ ввели в эксплуатацию высокопроизводительные вычислительные комплексы для основных оперативных технологий и научно-исследовательских разработок. Один из этих комплексов, установленный в Москве и самый мощный из трех, нам удалось посмотреть.

Длинные коридоры ГУ "ГВЦ Росгидромета" - что-то покрашено в оранжевый цвет, как здесь, что-то в зеленый

Отголоски старых времен, когда зарплата платилась наличными

Нам сюда - это кабинеты руководящего состава на "палубе"

Вообще, здание ГУ "ГВЦ Росгидромета" раритетное. Оно было построено в начале 60-х для огромной М20, которую разрабатывали в Институте точной механики и вычислительной техники (ИТМиВТ) и СКБ-245. Трансформаторы на местной подстанции, кстати, с того же времени выживают. Дальше в машинных залах стояли М-220, "Весна", далее серия БЭСМ (речь идет о моделях 4 и 6), линейки "Минсков" (начиная с модели 22, которая использовалась для обработки спутниковой информации), разные ЭС ЭВМ.

ГВЦ "Росгидромета" в то время котировался достаточно высоко, поскольку компьютеры попадали "с колес" в достаточно жесткие условия эксплуатации (т.н. "оперативный цикл"), что позволяло выявлять все их возможности до предела - разработчики были заинтересованы поставлять самые первые серийные образцы своего "железа". А в 1975 году было принято решение переориентироваться на американские ЭВМ, поскольку отставание отечественной промышленности было очень существенным, а с выходом на рынок IBM все наши направления ЭВМ были "задушены" и ГВЦ должен был получить машины CDC 7600. Правда, по политическим мотивам контракт распался и был поставлен только один CDC 7200. Но с тех пор в ГВЦ появились машинные залы, построенные по западным технологиям, далее были машины ЕС 1060/1066. И в 1996 году появился Cray VMP8E с мощностью 2,4 Гигафлопс, который отработал ровно 10 лет - благополучно скончался в 2006 году.

Фрагменты комплекса SGI Altix 4700 в ГУ "ГВЦ Росгидромета"

Фальшпол с системой водяного охлаждения

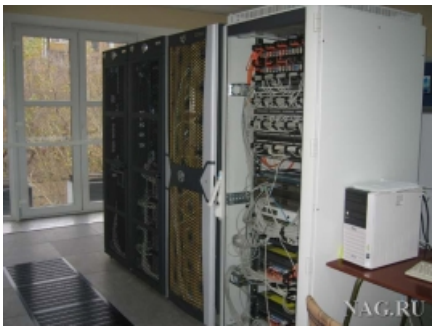
После этого, в 2009 году, в ГВЦ Росгидромета был смонтирован вычислительный комплекс с пиковой производительностью 27 TFlops (27 триллионов операций с плавающей запятой в секунду), который объединяет 2 кластера. Первый - кластер SGI Altix Ice 4700 с пиковой производительностью 11 TFLOPS (1664 процессорных ядра Intel Itanium2 9140M, оперативная память 6,6 Тбайт) с узлом в 128 процессорных ядер с единой оперативной памятью 512 Гбайт (4 Гбайт на процессорное ядро). Второй - кластер из стандартных серверов SGI Altix Ice 8200 с пиковой производительностью 16 TFLOPS (1408 процессорных ядра Intel Xeon e5440, оперативная память 2,8 Тбайт) с узлом из 8 процессорных ядер и оперативной памятью 16 Гбайт (2 Гбайта на процессорное ядро).

Электрический щит - "наследство" от суперкомпьютера Cray, вполне работоспособен.

Зал большой, но старый - инженерные коммуникации прокладывали после того, как из него же "вырвали" "скончавшийся" Cray

Сбоку на одном из кластеров примостилась Wi-Fi-точка Cisco

Нагрузка у кластеров по времени не отличается - оба работают 24 часа в сутки, однако тип этой загрузки различен. К примеру, Ice 4700 уже начинает выпускать оперативные прогнозы, работая с адаптированными для его мощности математическими моделями, в то время как Ice 8200 пока в большей степени загружен научными проектами - его возможности до конца еще не раскрыты. Но ресурсов Ice 4700 для оперативной работы уже недостаточно, поэтому на Ice 8200 будет мигрировать часть оперативной технологии. Интересно, что стойку с Ice 8200 во многом собрали из "остатков" контракта - дело в том, что в тендерном задании имелись в виду логические серверы, а сотрудники интегратора поняли, что речь идет о физических.



Комплекс G-Scale 4700 и его фрагменты в Новосибирском ЦГМС-РСМЦ

Комплекс G-Scale 4700 и его фрагменты в Хабаровском ЦГМС-РСМЦ

В машинных залах стандартная температура - 21 градус по Цельсию

Вычислительные комплексы Крафтвэй G-Scale S-4700 (104 процессорных ядра Intel Itanium2 9140M, оперативная память 208 Гбайт) в РСМЦ в г.г. Новосибирск и Хабаровск идентичны, имеют пиковую производительность 660 GFLOPS (660 миллиардов операций с плавающей запятой).
Большинство прогнозов в прогностических центрах готовят к 0 и 12 часам по Гринвичу, то есть два раза в сутки. Хотя, спецпотребители могут заказать себе и более частый прогноз по какому-то определенному географически значимому району. Росгидромет в настоящий момент дает уверенный прогноз на неделю вперед - с развитием суперкомпьютеров это временное "окно" возрастет вдвое. Оценка точности происходит не по одному, а по системе взаимосвязанных критериев, интегральная оценка, которую используют для простоты понимания - доля от 100%й вероятности. Разумеется, на каждое время она различна. На сутки вперед Росгидромет выдает прогноз с вероятностью 97-98%, на двое - примерно 92% и т.д. Отметим, что увеличение точности на ближайшие сутки на сотые доли процентов достигается очень большим трудом. С прогнозами на более длительное время ситуация несколько иная - сейчас точность на пятые сутки примерно аналогична той, что некоторое время назад была на трое суток.

С внедрением суперкомпьютеров увеличилась и детализация по площади. Мезомасштабные модели ранее считались на квадрате площади с горизонтальным разрешением 150 на 150 км, сейчас уже 7 на 7 км, а тестовые варианты краткосрочных прогнозов доступны на площади 3x3 км: именно для этого и нужны производительные кластеры, поскольку это не количественный, а качественный скачок в той физике, которая в них заложена.

Прообраз модели Ice 4700 - к сожалению, эта "железка" уже никому не нужна

В отдельном, "старом" зале находится много разнообразного оборудования. Часть систем - это те небольшие кластеры мощностью до 100 Гигафлопс, которые приобретались в период с 2006 по 2009 гг. для поддержания работы своих математических моделей. Некоторое "железо" осталось от тестовых стендов, которые собирались производителями оборудования для проверки на соответствие решаемым задачам ГВЦ Росгидромета - некоторые из них после подведения итогов тендера на поставку суперкомпьютера производители даже не стали забирать, поскольку техника уже устарела.

Рабочая лошадка - кластер на базе процессоров Itanium

Сбоку от кластера - блок управления, который используется для мониторинга или работы "в горячем" режиме.

Вытянутая стойка с юнитами - это первый кластер из четырех серверов, который был приобретен ГВЦ Росгидромета на базе процессоров Itanium. Это было в 2003 году и он работает до сих пор - решение было дополнено несколькими серверами на базе процессоров Xeon и нагружено для выполнения оперативных задач по "старым" математическим моделям. Кроме того, на будущее, как только на суперкомпьютерах Silicon Graphics полностью адаптируют новые модели расчета прогнозов погоды, старые кластеры нагрузят долгосрочными научно-исследовательскими проектами и вспомогательными задачами, на которые жалко отвлекать время производительного "железа". Все оборудование включено в локальную сеть - для того, чтобы дать ему задачу нет никакой необходимости приходить с CD или флэшкой.

Отдельные серваки для решения оперативных задач - все под нагрузкой

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Остатки Cray - железка так и не открылась, уж сколько лет в заклиненном состоянии :)

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Один из тестовых кластеров

Отдельные серверы, нагруженные выполнением отдельных прикладных задач - у них локальное управление, как единый кластер они функционировать не могут. Прогностические модели здесь не просчитываются, единственное, что может быть запущено - системы интерпретации.

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Операторский зал ГВЦ Росгидромета

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Рабочее помещение разделено на отдельные секторы - рабочие места разделяются по направлениям и используемым технологиям

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Слева от компьютера - раритетные системы связи

Операторский зал ГВЦ Росгидромета, в смену здесь работает четыре человека (всего в ГУ ГВЦ Росгидромета 95 сотрудников), для "присмотра" за автоматизированными системами, которые контролируются в режиме online, этого вполне хватает.

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Типичное рабочее место оператора Minimax - они обеспечивают прямую связь операторского зала с телекоммуникационным центром Росгидромета

[Кластеры Росгидромета: фотопрогулка по ВЦ]
На дисплеях - прогноз погоды, а вообще прогностическая ситуация поставляется в виде цифровых полей, на основании которых можно построить цифровые карты погоды на период от суток до двух недель вперед

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Наследство со старых времен

Здесь множество рабочих мест, на которых готовится необходимая для Росгидромета продукция по прогнозам погоды - возможно получать статистические данные и изображения, накладывать их друг на друга. Дело в том, что часть региональных прогностических центров на местах не занимается интерпретацией данных вообще, а получают готовые карты погоды для своей местности. В рамках ИТ-модернизации Федеральной службы по гидрометеорологии и мониторингу окружающей среды, подобное оборудование находится практически во всех 93 прогностических центрах, которые находятся на территории РФ.

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Система охлаждения в подвале

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Щит управления

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Резервуары для воды

Система охлаждения построена без единой точки отказа - любой элемент резервируется по принципу N+1. Она работает в двух режимах - или с помощью системы водяного охлаждения, второй вариант - забор холодного воздуха извне, это хорошо помогает экономить электричество зимой. Но это в теории - в последнее время зимы были настолько теплыми, что использовать этот режим особой необходимости не было. В цистернах хранится 14 тонн холодной воды.

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Новые UPS

От бросков напряжения суперкомпьютеры защищают два UPS (держат весь комплекс в течение 15 минут) - первый остался от старого Cray и новый, поставленный вместе с решением Silicon Graphics (две системы по 400 Кватт). Так вот на старом есть замечательный распределительный щит, который держит нагрузку и перебрасывает напряжение со стационарных вводов на резервные батареи. Его подключили, все заработало и... в один прекрасный день все выключилось. Оказалось, что подвел... светодиод за пару долларов, который показывает индикацию какая линия нагружена в настоящий момент. Лампочка из-за скачка напряжения буквально взорвалась и выключила систему. Многомиллионный комплекс выключился из-за мелочи.

[Кластеры Росгидромета: фотопрогулка по ВЦ]
Злополучный распределительный щит

Резервного генератора не предусмотрено по объективным причинам: установить такой объект в нескольких сотнях метров от здания Правительства РФ (т.н. "белый дом") практически нереально. Кроме того, после вывода из района Пресни множества заводов проблем с электропитанием не существует - двух независимых вводов вполне хватает, а никаких скачков из-за повышенного потребления уже несколько лет подряд нет (а раньше они были в районе 7 утра). Кроме того, если по какой-то причине выйдет из строя питание для суперкомпьютера в столице, работать с данными смогут центры в Новосибирске и Хабаровске, а также иностранные вычислительные центры - обмен информацией все равно не прекратиться.