Глава 4.
Поставка, монтаж и первичное тестирование кластера

Определившись с архитектурой кластерной системы, нужно решить еще один, казалось бы, формальный и не столь существенный вопрос: кто будет выполнять поставку кластерной системы? Предвидим, что первая реакция может быть примерно такой: железо – оно и есть железо, если цена устраивает, то какая разница, что за фирма нам его привезет? Не все так просто. Кроме стоимости есть немало вопросов, которые нужно принимать в расчет. Каковы сроки поставки? Входит ли в стоимость контракта монтаж кластера, его тестирование и настройка? Над этим стоит задуматься, если нет собственного опыта работы с подобного рода оборудованием или уверенности в качественном выполнении всех перечисленных этапов своими силами. И над этим стоит особо серьезно задуматься, если будут использоваться нестандартные устройства или комбинации компонентов. В одном из проектов Московского университета поставщик после оплаты и отгрузки нам коробок с оборудованием свое общение с нами фактически прекратил. Все наши настойчивые указания на то, что в материнских платах есть явный дефект, не позволяющий получать больше 30% от заявленной скорости передачи данных по интерфейсу SCI, оставались без внимания. Поставщик полностью устранился от решения проблем. То, что в такой ситуации принципиально снижается эффективность решения задач, для чего и создавался кластер с дорогой высокопроизводительной сетью SCI, в расчет не принималось. Осознав бессмысленность общения с этими людьми, мы стали работать с фирмой-производителем данных плат напрямую, которая к ее чести признала обнаруженный нами дефект собственной ошибкой и за свой счет заменила неисправное оборудование. На поиски проблем, осознание несостоятельности возлагавшихся изначально на поставщика надежд и

урегулирование всех вопросов потребовался примерно год. Естественно, что “услугами” того поставщика мы больше не пользовались, и всем советовали держаться от него подальше.

Влияет ли поставщик оборудования на качество решения прикладных задач? Как мы только что увидели, да, может повлиять самым непосредственным образом. Другая компания в недавнем проекте в конфигурацию кластера заложила интересную модель многовходового коммутатора InfiniBand: параметры превосходные, однако сильно смущало отсутствие практики ее использования. Решились, поскольку поставщик, будучи со своей стороны заинтересованным в освоении новой модели, взял решение потенциальных проблем на себя. Предчувствия не обманули, вопросы посыпались как из рога изобилия, но все они были решены на удивление быстро. Поставщик заранее установил необходимые контакты, быстро наладил линию общения с инженерами в режиме on-line, а затем для устранения найденной в оборудовании ошибки привез специалиста компании-производителя данного коммутатора на место установки кластерной системы. Все было сделано в кратчайшие сроки, никаких попыток переложить ответственность на кого-то еще не было. **Выбор надежного партнера по поставке оборудования** – это один из ключевых моментов, определяющих успех кластерного проекта в целом.

Других вопросов, о которых нужно также подумать при выборе поставщика, набирается немало. Входит ли в указанную стоимость гарантийное и сервисное обслуживание? Каковы сроки гарантийной замены вышедшего из строя оборудования? Выполняет ли поставщик настройку программного обеспечения или просто привозит ящики с аппаратурой? Компания может поставить только кластер или она в состоянии спроектировать весь вычислительный комплекс “под ключ”, включая системы хранения данных, энергоснабжения, климатического контроля, мониторинга, выполнить интеграцию всего комплекса в инфраструктуру организации? Немаловажный вопрос – наличие у компании опыта выполнения подобных проектов и наличие в штате

собственных квалифицированных инженеров. Не так сложно начать дело, гораздо сложнее постоянно поддерживать его выполнение на достойном уровне. С первичной настройкой поставщику смогут помочь сторонние специалисты, а кто поможет оперативно разобраться с нестандартными вопросами в последующие годы работы кластера? Хорошей проверкой основательности подхода к делу служит состав документации, которую поставщик дает вместе с кластерной системой: руководство оператора, администратора, пользователя, описание программно-аппаратной среды, параметров аппаратуры и настроек операционной системы, состав базового и специализированного программного обеспечения. Во всяком случае, при заключении контрактов и проведении конкурсных торгов ставьте это обязательным условием, тогда есть надежда получить, быть может, и не безупречный, но все же вариант документации.

А в целом, выстраивая общение с поставщиком, не забывайте время от времени задавать себе вопрос: “Что мне хочется больше: заниматься своими задачами или переквалифицироваться в системного программиста с инженерным уклоном?”. Действуйте далее, исходя из честного ответа.

Если принято решение собирать кластерное оборудование в комплекс самостоятельно, то продумайте и заранее опишите четкий план работы и последовательность всех

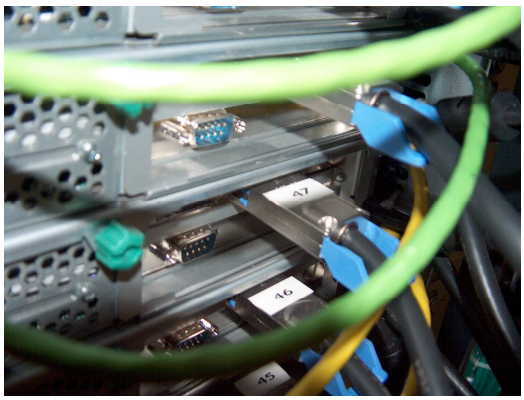

Рис. 4.1. Пример маркировки кабелей

действий. Начните с более общих вопросов и постепенно детализируйте их до тех пор, пока не получите для себя предельно ясную картину. Чтобы не запутаться в процессе сборки и настройки, используйте составленную на предыдущем этапе схему кластера, дополняя ее по ходу работ новой информацией.

Чтобы не запутаться в проводке, а ее будет много, проведите **предварительную маркировку кабелей** (рис. 4.1). Обратите внимание, что это необходимо сделать перед (!) коммутацией узлов. Сами узлы также полезно промаркировать. Для маркировки кабелей можно использовать обычные маркеры либо пластиковые стяжки с площадками и наклейки.

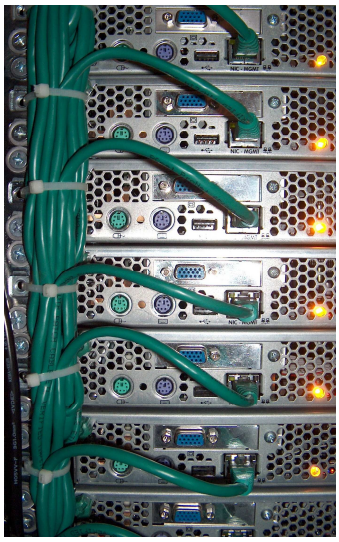

Рис. 4.2. Пример организации кабелей пластиковыми стяжками

Очень удобным может оказаться специальный маркировочный принтер (например, DYMO Letratag или аналогичный). Такой принтер печатает наклейки на самоклеющейся ленте и имеет размер чуть больше калькулятора. Полученные наклейки можно использовать как для маркировки кабелей (на площадках стяжек), так и для маркировки оборудования.

Большую помощь в креплении кабелей могут оказать **пластиковые стяжки**, не пренебрегайте ими (рис. 4.2). Это не только выглядит опрятно, но и намного удобнее в эксплуатации. Попробуйте разобраться с проблемами, которые возникли где-то в кабельном хозяйстве, показанном на рис. 4.3.

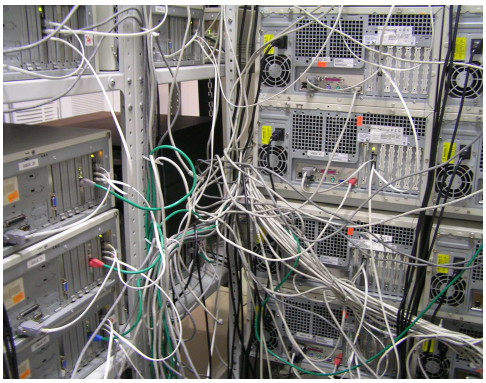

Рис. 4.3. Пример неудачной организации кабельного хозяйства

Установите и подключите источники бесперебойного питания. Как правило, подключение должен выполнять квалифицированный специалист. Обязательно тщательно **изучите инструкцию по установке и подключению UPS**. Ни в коем случае не подключайте UPS через сетевые фильтры ("пилоты" и т.п.), поскольку подобные фильтры крайне негативно влияют на работу источников. Не стоит включать UPS и в обычную бытовую сеть, так как частые перепады напряжения не прибавят им срока службы. Если инженерные возможности помещения позволяют, то оптимально включить их в специально выделенную компьютерную сеть через собственный автоматический выключатель.

После этого можно устанавливать вычислительные узлы и коммутаторы. Заранее продумайте, как будут проведены все кабели: питание, коммуникационная, транспортная и сервисная сети. Возможно, стоит заранее провести все или часть кабелей, а уже потом устанавливать узлы. Иногда некоторые проблемы при монтаже создают большие устройства, занимающие по 8-10 U и перекрывающие после своей установки доступ к некоторым гнездам, разъемам или же мешающие прокладке кабелей. Если такие устройства в конфигурации кластера есть, то их монтаж спланируйте особо тщательно, чтобы не пришлось подобные тяжести снимать и монтировать много раз.

После установки оборудования убедитесь, что источники бесперебойного питания находятся в рабочем состоянии. Подключите

питание ко всем компонентам. Теперь можно провести пробный запуск оборудования. Проследите, все ли узлы стартовали нормально.

Специально проверьте, что источники бесперебойного питания выполняют свою роль и в состоянии обеспечить нормальное функционирование необходимого оборудования при исчезновении питания. Это и в самом деле важный шаг, который сделать нужно именно сейчас до перевода кластера в режим эксплуатации. Реальный случай. В представительстве крупной европейской компании все критически важное оборудование подключили через мощный UPS, желая подстраховаться на случай возможного отключения питания. Такой случай возник один раз за пять лет работы, причем первым отключился тот самый UPS…

Для многих монтаж и интеграция кластера из россыпи оборудования в стойки является вполне выполнимым делом. Главное – это правильно оценить свои силы и возможности. И целесообразность. Стоит ли изучать особенности сборки, если даже не понятно, когда еще понадобится полученный в результате всего этого опыт. Может быть стоит обратиться за помощью к профессионалам, которые знают про подводные камни, да и гарантию дадут? Над этим выбором имеет смысл подумать.

Чтобы дать некоторое представление о трудоемкости сборки узлов кластера в стойку, опишем пример из нашей практики. Рассмотрим сборку кластера, состоящего из 80 узлов размера 1 U, четырех стандартных стоек с UPS на 12 кВт в каждой стойке, коммутатора InfiniBand, коммутатора Ethernet. В каждый узел нужно было дополнительно вставить второй процессор и сетевую карту InfiniBand, поставляемые отдельно от узлов.

Последовательность сборки можно описать следующим образом.

1. Установить в каждую стойку UPS и подключить блоки розеток.
2. Привинтить рельсы для всех узлов к стойке. Нами были заранее определены позиции узлов в стойке, чтобы в дальнейшем их не пришлось бы перемещать выше или ниже.
3. Выделить в комнате место для узлов, в которые уже вставлен процессор и плата InfiniBand.

4. Порядок операций с каждым узлом:
    - положить узел на ровную поверхность;
    - отвинтить крышку;
    - отвинтить заглушку разъёма для карты, вставить и привинтить карту InfiniBand;
    - отвинтить заглушку процессора, достать новый процессор, установить его в разъём, установить и привинтить радиатор;
    - завинтить крышку;
    - привинтить направляющие.
5. Установить узлы в стойки.
6. Установить коммутаторы в стойку.
7. Подключить узлы к UPS.
8. Соединить узлы сервисной сетью.
9. Подключить узлы к коммутатору Ethernet.
10. Подключить узлы к коммутатору InfiniBand.

Обратите внимание, что для каждого узла нам пришлось иметь дело с 19 винтами и целым множеством операций типа открутить/закрутить: 2 винта у крышки узла, 1 винт отвинтить от заглушки платы и завинтить его уже с платой, 2 винта открутить от заглушки процессора и завинтить их с радиатором, 8 винтов ушло на крепление рельсов в стойке, 4 – на крепление направляющих. Если фиксировать узлы в стойке, то нужно добавить ещё 2 винта. Столь детальный подсчет, возможно, и вызвал улыбку, однако из расчёта 80 узлов получим почти 2000 (!) операций завинчивания и отвинчивания. Советуем очень аккуратно сопоставить свои возможности с трудоемкостью всех операций по сборке кластера, что особенно актуально для больших конфигураций. Кластер СКИФ Cyberia – это 283 узла со всеми вытекающими отсюда последствиями для монтажа и сборки.

Операция установки процессора и радиатора очень тонкая, так как легко повредить процессор. Чуть криво поставленный радиатор при закреплении может просто расколоть керамическую основу процессора,

поэтому выполнять эту операцию нужно чрезвычайно аккуратно.

Несмотря на большую площадь помещения, при сборке неожиданно выяснилось, что очень трудно разместить в одном месте все имеющиеся коробки с оборудованием, уже собранные узлы, коробочки с процессорами и платами InfiniBand, а также пустые коробки и образующийся мусор. А расположить все это так, чтобы было еще и удобно работать – почти невозможно.

Для ускорения работы сначала во все узлы были установлены процессоры и платы InfiniBand, а затем все узлы были установлены в стойки. На каждый узел уходило от 15 до 25 минут, поэтому, даже работая вдвоём, на сборку всех узлов ушло более двух рабочих дней. Очень помогло использование аккумуляторных шуруповёртов. Узлы устанавливались вдвоём, для установки узлов в верхней части стоек не лишней оказалась стремянка. Монтаж каждого UPS смогли выполнить только втроём: использованные в нашем случае UPS HP R12000 RX вместе с батареями весили 220 кг каждый. Столь внушительный вес не является уникальной особенностью именно этой модели, источники бесперебойного питания всегда являются одной из самых тяжелых частей кластерной системы. Вес каждого устройства APC Smart-UPS 2200 VA RM в другом проекте составил почти 44 кг.

Схема проводки электричества и коммуникационных сетей была продумана заранее. Силовые электрические кабели и кабели управляющей сети укладывались вдоль противоположных боковых стенок стоек во время прикручивания рельсов в стойки.

Для подведения электричества от UPS к распределительному блоку потребовался кабель, способный длительно выдерживать ток в 50 А. Увы, выяснилось это только в процессе сборки кластера, так как устройство стойки не было изучено заранее. Да к тому же для аккуратной укладки купленного позднее кабеля нужного сечения пришлось открутить какое-то число уже установленных рельсов.

Перед покупкой кабеля рассчитывалось его сечение, для чего воспользовались данными из таблицы 4.1: для трехжильного (земля, ноль, фаза) медного кабеля получили значение 10 мм^2. Заметим, что для проводки внутри стоек необходимо пользоваться последними пятью колонками таблицы.

Питание к самим UPS было подведено заранее с привлечением специалиста. Подключение батарей и блоков логики в HP R12000 делается по строгой схеме, поэтому нельзя просто вставить все блоки внутрь и подключить нагрузку. После установки батарей, была запущена программа самотестирования, которая работала несколько десятков минут. До окончания ее работы никакой нагрузки к UPS подключать нельзя. Читайте инструкции!

Сечение жилы, мм^2	Ток, А, для проводов, проложенных					
	открыто	в одной трубе				
		двух одно-жильных	трех одно-жильных	четырех одно-жильных	одного двух-жильного	одного трех-жильного
1	17	16	15	14	15	14
1,2	20	18	16	15	16	14,5
1,5	23	19	17	16	18	15
2	26	24	22	20	23	19
2,5	30	27	25	25	25	21
3	34	32	28	26	28	24
4	41	38	35	30	32	27
5	46	42	39	34	37	31
6	50	46	42	40	40	34
8	62	54	51	46	48	43
10	**80**	**70**	**60**	**50**	**55**	**50**
16	100	85	80	75	80	70
25	140	115	100	90	100	85
35	170	135	125	115	125	100
50	215	185	170	150	160	135
70	270	225	210	185	195	175
95	330	275	255	225	245	215
120	385	315	290	260	295	250
150	440	360	330	–	–	–

Табл. 4.1. Допустимый длительный ток для проводов и шнуров с медными жилами с резиновой или поливинилхлоридной изоляцией

После узлов в стойки устанавливались коммутаторы InfiniBand и Ethernet. Всё оборудование сразу подключалось к работающим UPS, так как кабели питания были проложены заранее. Далее были проложены кабели InfiniBand и транспортной сети Ethernet.

В каждый узел приходит четыре кабеля: транспортный и сервисный Ethernet, InfiniBand, электропитание. Это означает, что всего было проложено более 320 кабелей, коммутировано более 640 разъёмов. Чтобы элементарно не запутаться в таком количестве разъёмов и кабелей, использовались Ethernet-кабели различного цвета, а также маркировка разъёмов. Для организации кабелей пришлось использовать более 200 пластиковых стяжек.

Такова реальная последовательность шагов, такова примерная трудоемкость. Хотите ли делать это самостоятельно, можете ли сделать это самостоятельно, целесообразно ли делать это самостоятельно или все же имеет смысл обратиться к профессионалам – решайте сами, оценив особенности своего проекта, свои технические, финансовые, кадровые и временны́е ресурсы. Оценивайте не абстрактно, а с точки зрения будущего использования кластера как инструмента решения конкретных задач.

Итак, оборудование собрано, и осталось его оживить. Без необходимого программного обеспечения это всего лишь красивый и внушительный монумент, однако, именно на этой основе будет создаваться полноценный, удобный и функциональный инструмент – вычислительный кластер. В последующих разделах данной работы мы обсудим вопросы установки и настройки операционной системы, систем и сред параллельного программирования, тестирование кластерной системы и определение характеристик ее работы, состав вспомогательных инструментов, вопросы использования специализированных прикладных пакетов.