



Basics of Cisco Optimized Edge Routing (OER)

Peter J. Welcher

Introduction

This article is being written at the beginning of 2006. I've had some time off during which I could experiment in the lab with a relatively new Cisco feature, Optimized Edge Routing (OER). And that's the subject of this month's article.

We'll look shortly at why you might want Optimized Edge Routing (OER) in your network. The short answer is that OER allows you to intelligently choose between multiple Service Providers (or WAN Providers) in how you route traffic outbound from a site. OER is a fairly big topic, so this article will have to focus on the Big Picture (what is OER, what does it do, how can I use it). A later article will then cover basic OER configuration. There are a lot of ways you can configure OER, so we may never cover all the "bells and whistles".

While explaining OER, I'll assume you already know something about NetFlow and IP SLA (IPSLA), formerly Service Assurance Agent. OER monitoring makes use of one or both of these services that are built into Cisco IOS code. In case you don't already know about these technologies, [links](#) to my previous articles on those topics are at the end of this article.

What Is OER?

Optimized Edge Routing (OER) is intended for sites using multiple Internet or WAN Service Providers. It seems like you could also use it on CE routers with dual MPLS VPN connections.

The intent of OER is to **automatically** detect Internet (or IP-based WAN / MPLS VPN Service Provider) "blackouts and brown outs", that is total or partial packet loss or service degradation. Once service degradation is detected, OER then **automatically** responds by routing traffic around the problem. This differs from normal routing, which is focused on detecting a routing path, rather than the condition of the service over that path.

Cisco OER can respond to policy violations concerning:

- 1 response time
- 1 packet loss
- 1 path availability
- 1 traffic load distribution

OER can also be used to optimize traffic load distribution dynamically in response to load, including monetary cost minimization on links. To prevent instability, traffic shifts cannot be instantaneous, but are generally in response to conditions lasting one or more minutes.

OER uses a router or appliance functioning as a master controller. It communicates with one or more border routers that connect to the ISP or WAN providers. The controller is generally placed in proximity to the border routers it controls. The links to the controller should be reliable and high-speed for proper OER operation.

Where Can I Run OER?

OER was introduced in 12.3(8) T code, with features added in 12.3(11) T, 12.3(14) T, and 12.4(2) T. I've been running it in a Cisco 1841 router running 12.4(3) (no T). You should check Feature Navigator for which hardware and software supports OER. I'd expect the answer to be newer routers with a fairly solid amount of memory and capable CPU.

Cisco OER requires:

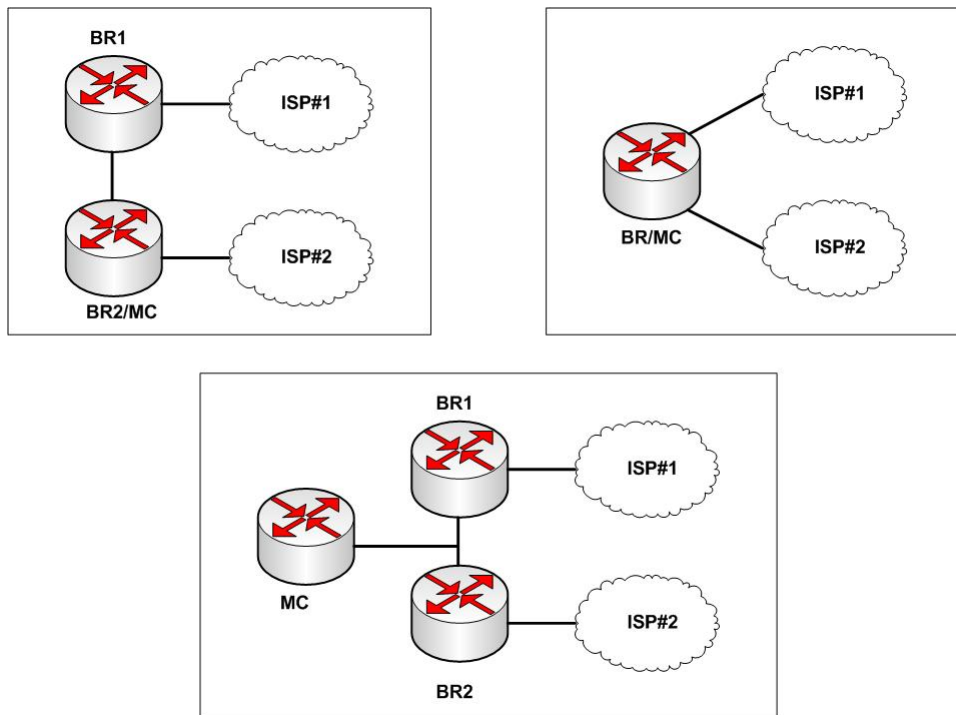
- 1 Enabling Cisco Express Forwarding (CEF) on participating routers
- 1 Establishing consistent routing prior to OER deployment
- 1 Redistributing static routes into your Internal Gateway Protocol (IGP). A tag is used to facilitate redistribution controls.

Restrictions on OER:

- 1 OER does not affect inter-domain routing.
- 1 OER does not affect interfaces that are not under OER control.
- 1 OER does not influence asymmetrical routing.
- 1 OER monitors and controls **outbound** traffic only.
- 1 For VPNs, OER supports only IPsec/GRE VPNs.
- 1 Token Ring interfaces are not supported by OER.

OER Design Topologies

The following figures show some representative OER border router and master controller router topologies.



Other topologies can be made to work. I would claim the above are good topologies since they will make it easier to understand and troubleshoot OER behavior in.

Caution: OER Border Routers must use outbound next hops that are on different subnets. Consequently, Internet exchange points where the border router communicates with several service providers over the same broadcast network are not supported.

The basic requirement is that the OER managed network must have border routers with at least two external (ISP connected) interfaces between them, and at least one internal interface connecting to the inside network per border router. When you configure OER, you tell the master controller about each border router and its internal and external interfaces. You also set up authentication between them.

Any internal interfaces you specify are used only for passive monitoring with NetFlow. If you do use passive monitoring, you do not have to explicitly configure NetFlow.

OER also uses the concept of a local interface, which is the source for communications from border router to master controller. A loopback interface might well be used for this. If a router is running as both master and border, then a loopback interface should be used as the local interface.

The master controller router does not need to be in the traffic forwarding path. It does have to be reachable by the border routers. OER can support up to 10 border routers, and 20 OER managed external interfaces. See the link to the performance document below to figure out how to size a router appropriately for your needs.

How Does OER Work?

A high-level description of how OER works is appropriate at this point. We can then drill down on selected details later.

OER configuration starts by telling the master controller and border routers how to communicate with each other, and which interfaces are internal and external.

Once that very basic OER setup is in place, OER must either be configured to automatically learn IP prefixes to track, or else it must be manually told which IP prefixes to track .

Each border router monitors information about each prefix and performance statistics over each external interface. This information is periodically reported to the master controller. If the prefixes and exit links comply with configured policy, routing is left as is. You can specify different policies for different groups of prefixes or for different exit links.

If a prefix or link does not comply with the configured policy, a policy-based decision is made by the controller. The border routers then may be commanded to alter routing to bring the prefix or exit link into compliance with policy. This is done one prefix at a time.

OER routing control is exerted by injecting routes into the border routers. This is done via OER command messages from the master controller to the border routers, and not by inserting routes on the master controller. Currently, OER can influence routing in two ways:

- 1 Setting the BGP local preference for a specific prefix
- 1 Creating a temporary static route for a specific prefix

It does this only for prefixes with a some superset prefix present in the routing table, to avoid routing loops. (Default route plus "ip classless" acts as a superset route prefix for all more specific prefixes.) The BGP routing table is searched first for a parent route. If one is found, a more specific BGP route is added. If not, the static routing table is searched for a parent route. If one is found, a more specific static route is temporarily added to the routing table.

This routing change at the border routers influences the other routers in the internal network through one of the following methods:

- 1 Internal BGP peering
- 1 BGP or static route redistribution into the IGP

Concerning this last point, if you have border routers in close proximity (namely, with a high speed LAN connection between them), you can use default routes to get packets to the border, and then have OER shift some traffic for selected prefixes between the two exit routers. OER is mainly about preferring one border router to the other. The IGP routing only comes into this if you have to rely on your IGP to route traffic between the border routers, or if you want optimal routing, directly to the "correct" border router. Personally, if I've got more specific routes being added and removed on my border routers, I would prefer for the IGP to see as little of that churning activity as possible. Exactly how to handle this is one design choice.

The injected BGP or static route is not advertised to external peers, and has no routing impact outside the local site.

By the way, there are plenty of show commands allow you to monitor what OER is doing. Syslog log messages also give a solid indication of decisions made by OER.

OER Prefixes

OER is based on monitoring performance for selected prefixes. When you configure OER, you have to tell it which prefixes to monitor and what aspects of performance to monitor.

You can configure the border routers to automatically learn prefixes based on NetFlow Top Talker statistics. When you configure this aspect of OER, the necessary NetFlow features are configured on the router. The top 100 talkers (flows) are learned by default. You can configure the router to learn up to 5000 flows this way. The learned prefixes can be aggregated in several ways. The default is /24. In the lab, I observed flows to several hosts on the same subnet, which aggregated to one /24 prefix. Other aggregation methods can be configured.

You can also manually specify prefixes that are important to you. This is done using an oer-map and an ip prefix-list.

As of 12.3(11) T you can specify learning based on IP protocol or TCP/UDP port numbers or ranges of ports. This can restrict learning to only those applications that are of interest, reducing CPU and memory impact on the router.

It is a good idea to bear in mind that all learned prefixes are being monitored, generally infrequently. We will discuss monitoring next, but for now bear in mind that not only does the router have to periodically check measurements, but it may be generating traffic to produce those measurements. The more prefixes learned, the more work the router has to do.

Prefixes pass through states after they are learned. The states are as follows:

- 1 Default: Not under OER control, but routed based on existing routing. Prefixes start out in this state.
- 1 In-Policy: The status of the prefix matches default or configured policies. No changes are made when in this state until the configuration or performance measurements change.
- 1 Out-of-Policy: The prefix does not match policy. Active probing or passive monitoring (or both) will be used to find a better exit, while the prefix is in this state. If all exit links fail policy (are out-of-policy), the master controller uses the best one.
- 1 Choose: The master controller is choosing an exit link. (Don't blink or you may miss this state?)
- 1 Holddown: The master controller moved the prefix to a new exit. No policy changes are applied while the prefix is in holddown state. This is intended to prevent flapping.

Overly aggressive policy settings can cause a prefix or exit link to remain in the out-of-policy state. (There's no pleasing some people!)

There's a point to the above state information. The only way OER can get data about an alternative path is to alter the routing for a prefix and gather more data. Holddown state is an indication that it is doing so. If you are working with OER in the lab, be aware that it may take a few minutes for OER to check out the alternative paths for each prefix.

OER Monitoring

OER monitors the set of prefixes either via active probing or passive monitoring, or both.

Passive monitoring amounts to looking up NetFlow data in memory. That means the router observes what happens when packets are sent, and record the results as internal NetFlow statistics. If there are no packets being sent, there is no new data for the system. NetFlow data captures delay and throughput statistics. The delay measurements are based on TCP RTT (initial SYN to following ACK). The OER data also records Packet Loss (observing highest TCP sequence number, and received packets with lower sequence number) and Unreachables (SYN with no received ACK) for passive measurements.

The documentation notes "OER passive monitoring is based on TCP traffic flows for IP traffic. Passive monitoring of non-TCP sessions is not supported." I presume this is because with UDP you cannot readily obtain delay estimates, response counts, etc.

Active probing defaults to ICMP echo, ping. When I activated active monitoring, I turned on debugging (**debug ip icmp**). I did see periodic pings to various destinations. Note that repeated ping probing might trigger an IDS or IPS on the remote site.

OER active probing can be configured to use IP SLA (formerly SAA) measurements instead of ping. This allows OER to

respond to delay or jitter in the network. Currently, OER can use ICMP, TCP connections, or UDP echo. Note that the target for the latter two must be capable of responding. If the target is a router, it must be configured with "**rtr responder**".

As of 12.3(14) T, OER can do traceroute probes. These collect delay, loss, and reachability information for each hop from source address to probe target prefix. You can configure such probes to run in three ways: continuous (run all the time), policy based (run only when the prefix is out of policy), or on-demand (run when you use the command "**show oer master prefix <prefix>/<length> traceroute**" with the **current** or **now** keywords and a specific prefix).

Monitoring tracks short term (5 minute) and long term (60 minute) measurements. These are compared for relative thresholding (i.e. when your policy is responding to changes compared to long-term behavior). Your policy can also threshold based on absolute measurements. Statistics you can monitor include:

- 1 lowest delay
- 1 highest outbound throughput
- 1 relative or absolute packet loss
- 1 relative or absolute link cost
- 1 prefix reachability

Timers

There are several timers that can affect OER behavior. OER has to first learn a prefix. The defaults cause this to take quite some time initially. Once prefixes are learned, monitoring has to collect statistics for the router to discover that a prefix is out-of-policy. How fast that will happen depends on how often measurements are made, and how often the router checks those measurements for each prefix. If the current route is out -of-policy, the router then may have to try another path, to see if it is better. That is where the holddown timer applies.

For learning prefixes, the following timers apply:

- 1 The **periodic-interval** command governs how often the master controller learns new prefixes from the border routers,. default 120 minutes. It is specified in minutes. If you don't make this number something like 3 minutes in your lab, you won't see much learning occurring! You will probably lose patience first.
- 1 The **monitor-period** command governs how long the top talkers are observed for, default being 5 minutes. This timer is also specified in minutes. So by default, every 2 hours (120 minutes) the border routers are sampled for 5 minutes to obtain top talkers. For a lab, you might want 3 minutes and 1 minute.
- 1 The **expire after** command governs how long before learned prefixes are expired (removed from the cache). Note that new prefixes are not learned if more than 90% of memory is in use. When memory is needed, inactive prefixes are removed, oldest first.

Monitoring tracks statistics over 5 and 60 minute periods. The 60 minute data is rolled up from the 5 minute measurements. Passive monitoring just accumulates statistics in NetFlow entries, so there is no specific timing of measurements. With passive monitoring, the only relevant timing is how often the controller checks the measurements. Active probes are sent once per minute (and this is apparently not configurable). If you are using traceroute probes, the probe interval is configured with the **traceroute probe-delay** command, default 1 second (1000 milliseconds).

The **periodic** command governs how often the master controller checks each prefix for policy compliance, default 5 minutes (300 seconds). The show commands I used showed "periodic 0", but the configuration showed no entry, that is, the default setting of 5 minutes.

The **holddown** timer defaults to 5 minutes (300 seconds) for dampening. This can be changed by configuration.

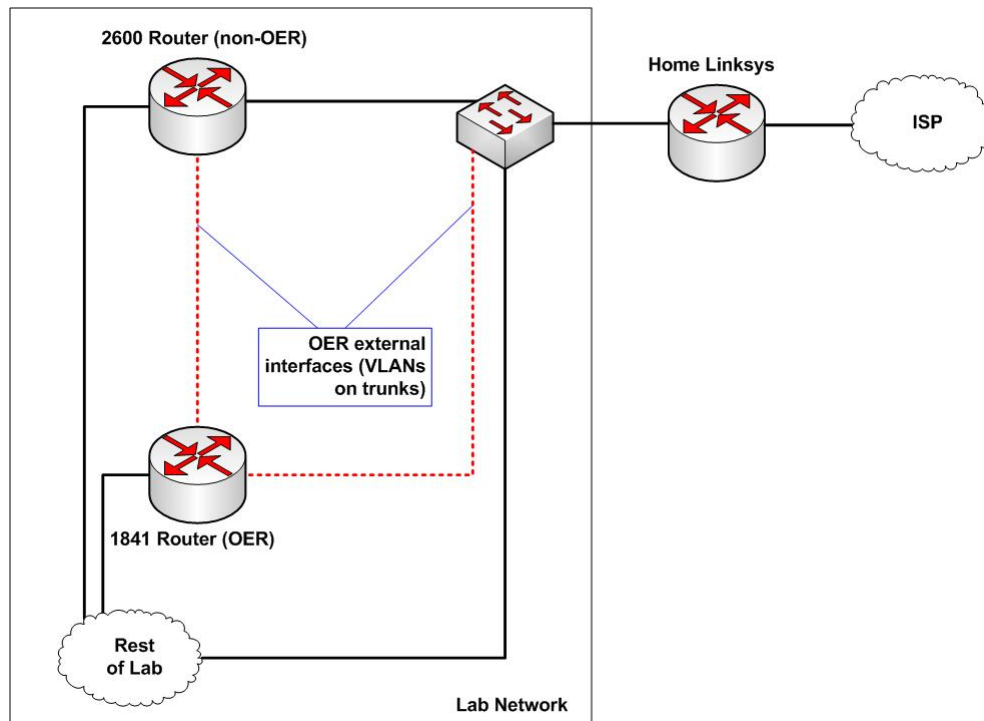
So in a steady-state situation, the probe and periodic timers would be the dominant timing factors. If the key prefixes keep changing, or behavior via another interface/peer has to be measured, then the holddown timer may also come into play. Learning timers may also be a factor increasing failover time. In short, OER has complex timing behavior defying verbal summarization beyond this!

Lab Observations

Although it is a mild sidetrack, I'd like to allow some reality to intrude.

In my lab, my two OER router outbound interfaces were selected vlan subinterfaces on two trunks. One went to the LAN

that connected my lab exit points to one of my home Linksys routers and the Internet. The other went to another lab router (2600) with a connection to the home LAN. See the following diagram.



[If you're wondering why I use a Linksys, bear in mind that I or others might be working with the lab gear, either locally or via outbound IPsec connection(s) from lab routers to a central VPN device. One strong requirement is that Internet access for my wife and children not be too easily messed up by anything configured in the lab devices. That means using a separate router that is not part of the lab. Linksys was the selected cost-effective alternative.

Other lab info: the lab routers both have static default routes pointing at the Linksys' inside interface (192.168.1.1). These are redistributed into EIGRP. Both do PAT of inside lab addresses to the address of their external interface (192.168.3 and 4, respectively). The Linksys does PAT again, to whatever address it picked up from Comcast (my ISP) via DHCP. The lab 2600 acts as DHCP server for the inside lab network. Lab computers can readily access the Internet, for example to do Windows Update.

I can wirelessly connect to the Linksys or to a WAP inside the lab network, so I made sure I was connected to the lab network (with appropriate static routes on my PC). I also made sure my PC was set with only default gateway being the OER router, since normally my lab DHCP supplies both router inside addresses as default gateways to the PC.

I then tested OER failover. I used several web browser windows to observe this. (www.netcraftsmen.net, www.cnn.com, www.wjla.com, www.abc.com).

To cause a failure, I disallowed the external VLAN on the first trunk on the switch the lab and Linksys connect to, the one from the OER router to the switch. This caused packet black-holing while leaving the interface up. Since I was using static default routing, there was no way for normal routing to discover the outage. If you want to get picky, the ARP entry would eventually have timed out.

Gotcha: If you are only doing OER on one router, as I was, and if you're using static default to the outside world, when OER failover kicks in, you'll have a routing loop: the OER temporary static route sends traffic to non-OER router. The static route represents a more specific prefix. When redistributed into EIGRP it causes the non-OER router to send packet back to OER router. Presto, loop! This is one case where you do NOT want to follow the OER configuration directions blindly.

When I did this failover testing, I noticed as expected that after I induced the packet black holing, the web pages would not refresh for a while, until OER re-routed the outbound traffic. I eventually saw most OER prefixes shift correctly to the other interface, the one that still worked. Some prefixes did not shift. I believe this was due to changes of host addresses, for example with www.cnn.com. I suspect these sites DNS load balance so that at different times the hostname resolves to different addresses. In particular, CNN appears to be using several /24 blocks, so when DNS name resolution switched to

a new block, OER would have to go through its multi-minute learning and holddown cycle all over again. Specifying this prefix or an aggregate CNN prefix manually would presumably have helped with this (not tried due to time constraints).

I then shifted to active monitoring, to see if failover went faster. My sense is that it did go faster, but I'd be hard pressed to quantify that. OER appears to distribute the workload somewhat, causing it to determine that different prefixes are out of policy at different times. I did note that the prefix(es) for `www.cnn.com` were still not shifting to the working link. I then tried to ping the specific host addresses visible in output from the command "**show oer master prefix detail**" for those prefixes. They did not respond to ping. My suspicion is that active monitoring was not seeing any difference between the two links (no ping replies with either), so both were out of policy. Apparently the one that was used formerly was somehow "better", even though down.

Policy

Your configured policy tells the router how often to do things, which statistics you care about, and how to respond to out-of-policy conditions.

A global policy specifies default timers and measurement thresholds for OER. Exit link policies are applied when you configure border router exit interfaces. These can include fiscal cost minimization formulas.

An OER map is used to specify a detailed policy for prefixes. An OER map is similar to a route map. It contains blocks with match and set clauses. You may only have one match clause per each oer map sequence block. The match condition specifies which prefixes the policy block applies to. The set clauses specify the policy for the matching prefixes. You are allowed to match learned addresses, so that you do not need to know those prefixes in advance.

For each prefix, the OER map allows you to specify timers (backoff, holddown) and out of policy thresholds for measurements (that is, when to react to measurements).

Global resolve priorities can be configured when there are multiple policies that apply. By default, delay has highest priority, and utilization second highest. Cost, loss, and utilization range can also be prioritized. A variance can also be specified, so that traffic is not re-routed when measurements are fairly close.

Modes

OER can be run in observe mode or in control mode. In observe mode, OER tells you what it would do, via show commands and syslog messages, but it does not actually alter any routing. In control mode, OER actually does alter routing.

My initial reaction when I tried this was that observe mode seemed to be working but I couldn't figure out what it was telling me. So I recommend some lab time to get used to how OER works and what the show commands tell you. You might then consider **cautiously** trying it live on the production border routers in observe mode. Make sure they have plenty of memory and CPU. And even then, you might first try it during a maintenance window, just in case of unexpected ill consequences. After all, you don't want to become known as "the person who brought down the Internet link the other day". Or worse yet, "the person who is now looking for a new job".

Summary

I promised links for prior articles about NetFlow and IP SLA (SAA). The links follow. The prior articles summarize NetFlow and IP SLA (SAA). They also contain links to key Cisco documents on these topics.

NetFlow and IPFIX	http://www.netcraftsmen.net/welcher/papers/netflow02.html
NetFlow	http://www.netcraftsmen.net/welcher/papers/netflow.html
Service Assurance Agent (SAA) and the Management Engine	http://www.netcraftsmen.net/welcher/papers/saa.html

Here are some Cisco Optimized Edge Routing (OER) links that I found useful in preparing this article:

OER	http://www.cisco.com/en/US/partner/products/ps6628/products_ios_protocol_option_home.htm
-----	---

Introduction/Main
Page

Cisco OER White Papers http://www.cisco.com/en/US/partner/netsol/ns471/networking_solutions_white_papers_list.htm

Primary and Backup WAN Links Using Cisco Optimized Edge Routing http://www.cisco.com/en/US/partner/netsol/ns471/networking_solutions_white_paper0900aecc

Cisco Optimized Edge Routing Overview http://www.cisco.com/application/pdf/en/us/guest/products/ps5870/c1161/cdccont_0900aecd8f

Cisco Optimized Edge Routing Deployment Guide http://www.cisco.com/en/US/partner/netsol/ns471/networking_solutions_white_paper09186a0c

OER Performance Tests http://www.cisco.com/application/pdf/en/us/guest/netsol/ns471/c664/cdccont_0900aecd80270c

Cisco IOS OER Configuration Guide http://www.cisco.com/univercd/cc/td/doc/product/software/ios124/124tcg/toer_c/ht_oer.htm

Cisco IOS OER Command Reference http://www.cisco.com/univercd/cc/td/doc/product/software/ios124/124tcr/toer_r/index.htm

OER Data Sheet http://www.cisco.com/en/US/partner/products/ps6599/products_data_sheet0900aecd801dfcec

Your comments, questions, and suggestions for future articles are of course welcome! See below to decipher my email address.

Dr. Peter J. Welcher (CCIE #1773, CCSI #94014, CCIP) is a Senior Consultant with Chesapeake NetCraftsmen. NetCraftsmen is a high-end consulting firm and Cisco Premier Partner dedicated to quality consulting and knowledge transfer. NetCraftsmen has ten CCIE's, with expertise including large network high-availability routing/switching and design, VoIP, QoS, MPLS, IPSec VPN, wireless LAN and bridging, network management, security, IP multicast, and other areas. See <http://www.netcraftsmen.net> for more information about NetCraftsmen. Pete's links start at <http://www.netcraftsmen.net/welcher>. New articles will be posted under the Articles link. Questions, suggestions for articles, etc. can be sent to **[pjw <at> netcraftsmen <dot> net](mailto:pjw@netcraftsmen.net)** (formatted this way to fool email harvesting software).

1/11/2006

Copyright (C) 2006 Peter J. Welcher