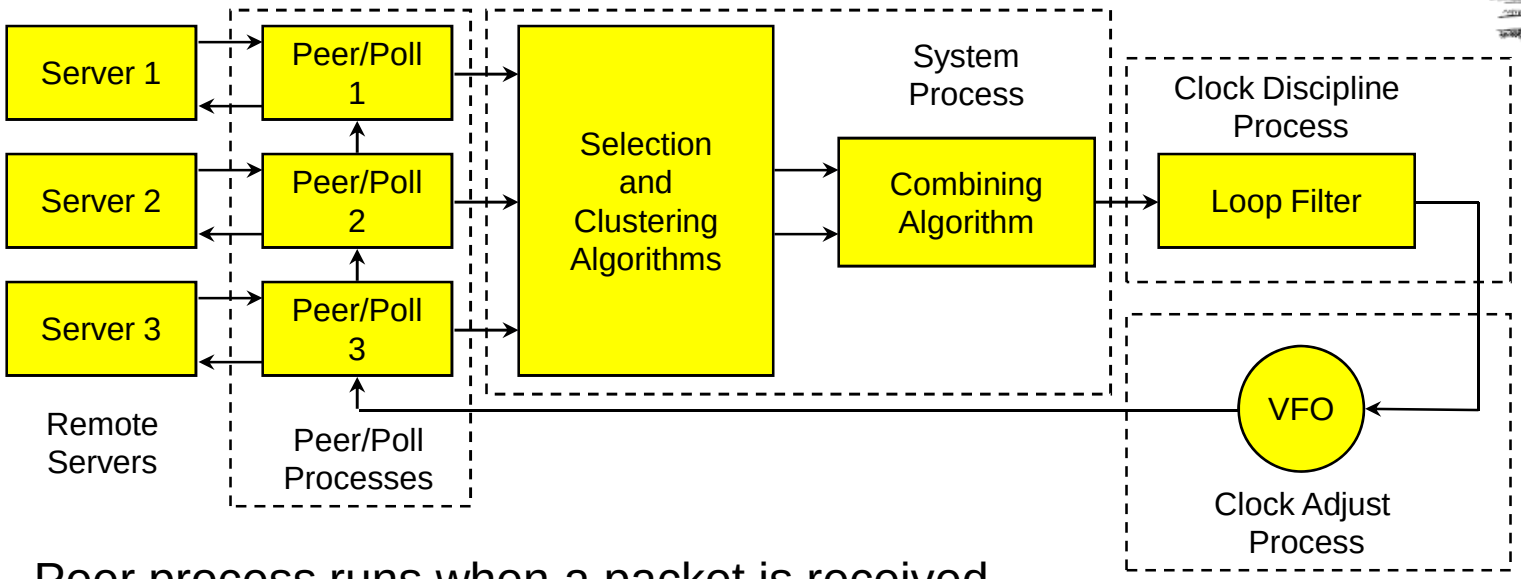
NTP Architecture, Protocol and Algorithms

David L. Mills
University of Delaware
http://www.eecis.udel.edu/~mills
mailto:mills@udel.edu

Sir John Tenniel; *Alice's Adventures in Wonderland,* Lewis Carroll

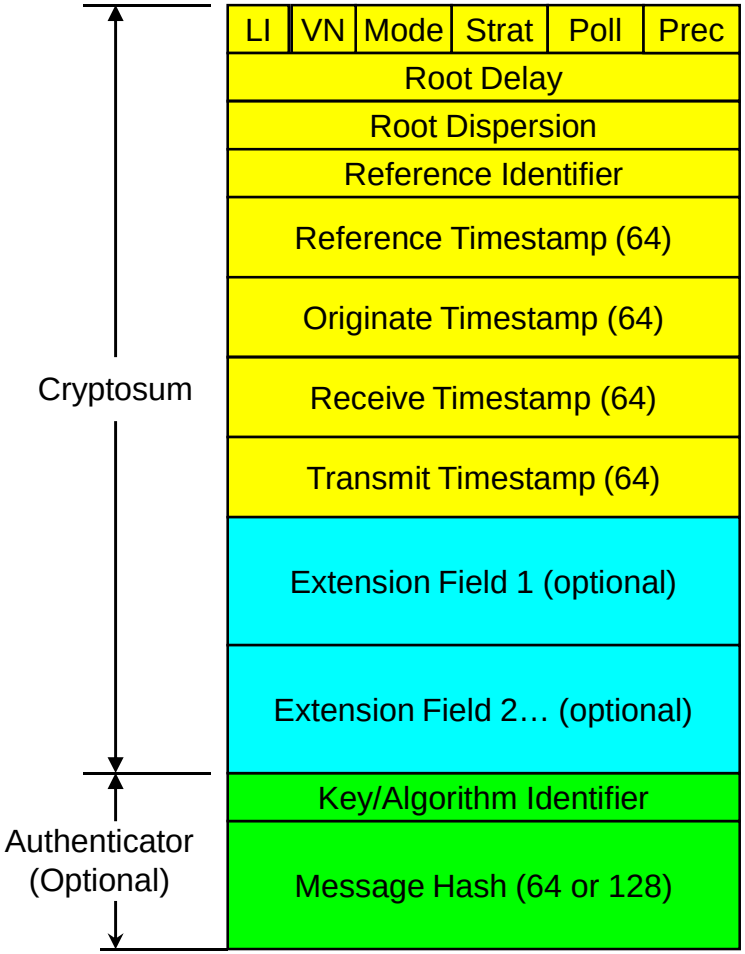
Process decomposition



- Peer process runs when a packet is received.
- Poll process sends packets at intervals determined by the clock discipline process and remote server.
- System process runs when a new peer process update is received.
- Clock discipline process runs at intervals determined by the measured network phase jitter and clock oscillator (VFO) frequency wander.
- Clock adjust process runs at intervals of one second.

NTP protocol header and timestamp formats

NTP Protocol Header Format (32 bits)

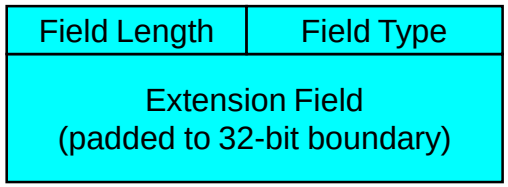


LI	leap warning indicator
VN	version number (4)
Strat	stratum (0-15)
Poll	poll interval (log2)
Prec	precision (log2)

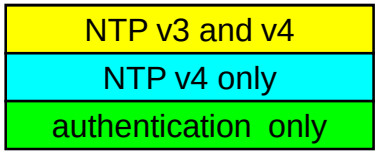
NTP Timestamp Format (64 bits)

Value is in seconds and fraction since 0^h 1 January 1900

NTPv4 Extension Field



Last field padded to 64-bit boundary

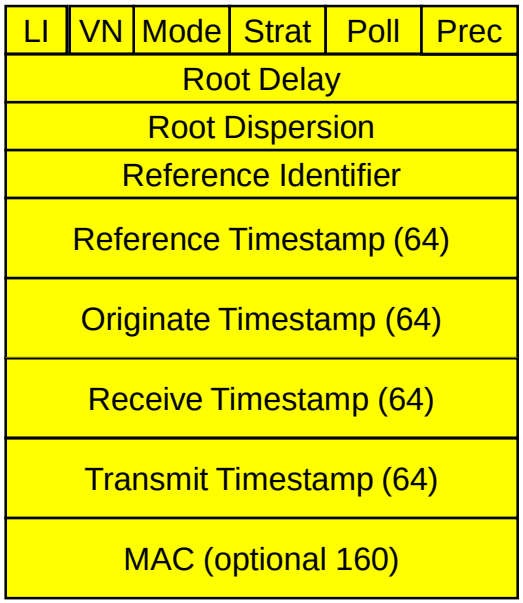


Authenticator uses DES-CBC or MD5 cryptosum of NTP header plus extension fields (NTPv4)

NTP packet header format

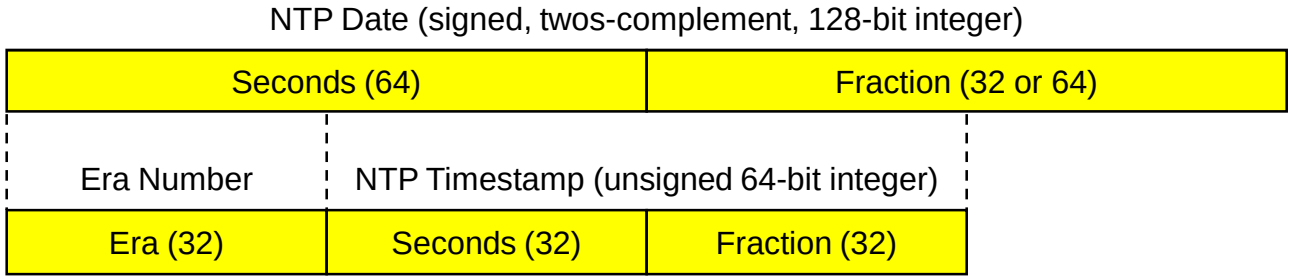
Packet header Variables	Description
leap	leap indicator (LI)
version	version number (VN)
mode	protocol mode
stratum	stratum
τ	poll interval ($\log_2$ s)
ρ	clock reading precision ($\log_2$ s)
Δ	root delay
E	root dispersion
refid	reference ID
reftime	reference timestamp
T_1	originate timestamp
T_2	receive timestamp
T_3	transmit timestamp
T_4	destination timestamp*
MAC	MD5 message hash (optional)

* Strictly speaking, T_4 is not a packet variable; it is the value of the system clock upon arrival.

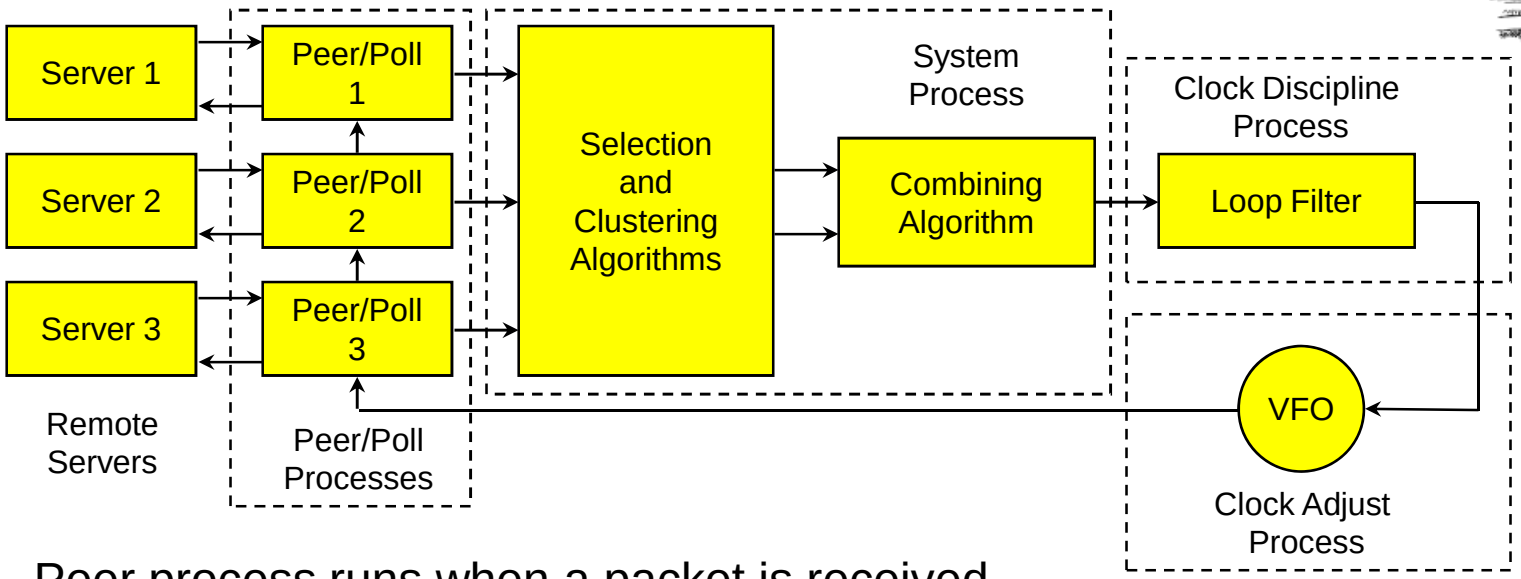


NTP date and timestamp formats and important dates

Year	M	D	JDN	NTP Date	Era	Timestamp		
-4712	1	1	0	-208,657,814,400	-49	1,795,583,104	First day Julian Era	
1	1	1	1,721,426	-59,926,608,000	-14	202,934,144	First day Common Era	
1582	10	15	2,299,161	-10,010,304,000	-3	2,874,597,888	First day Gregorian Era	
1900	1	1	2,415,021	0	0	0	First day NTP Era 0	
1970	1	1	2,440,588	2,208,988,800	0	2,208,988,800	First day Unix Era	
1972	1	1	2,441,318	2,272,060,800	0	2,272,060,800	First day UTC	
2000	1	1	2,451,545	3,155,673,600	0	3,155,673,600	First day 21st century	
2036	2	7	2,464,731	4,294,944,000	0	4,294,944,000	Last day NTP Era 0	
2036	2	8	2,464,732	4,295,030,400	1	63,104	First day NTP Era 1	
3000	1	1	2,816,788	34,712,668,800	8	352,930,432	4294967296	

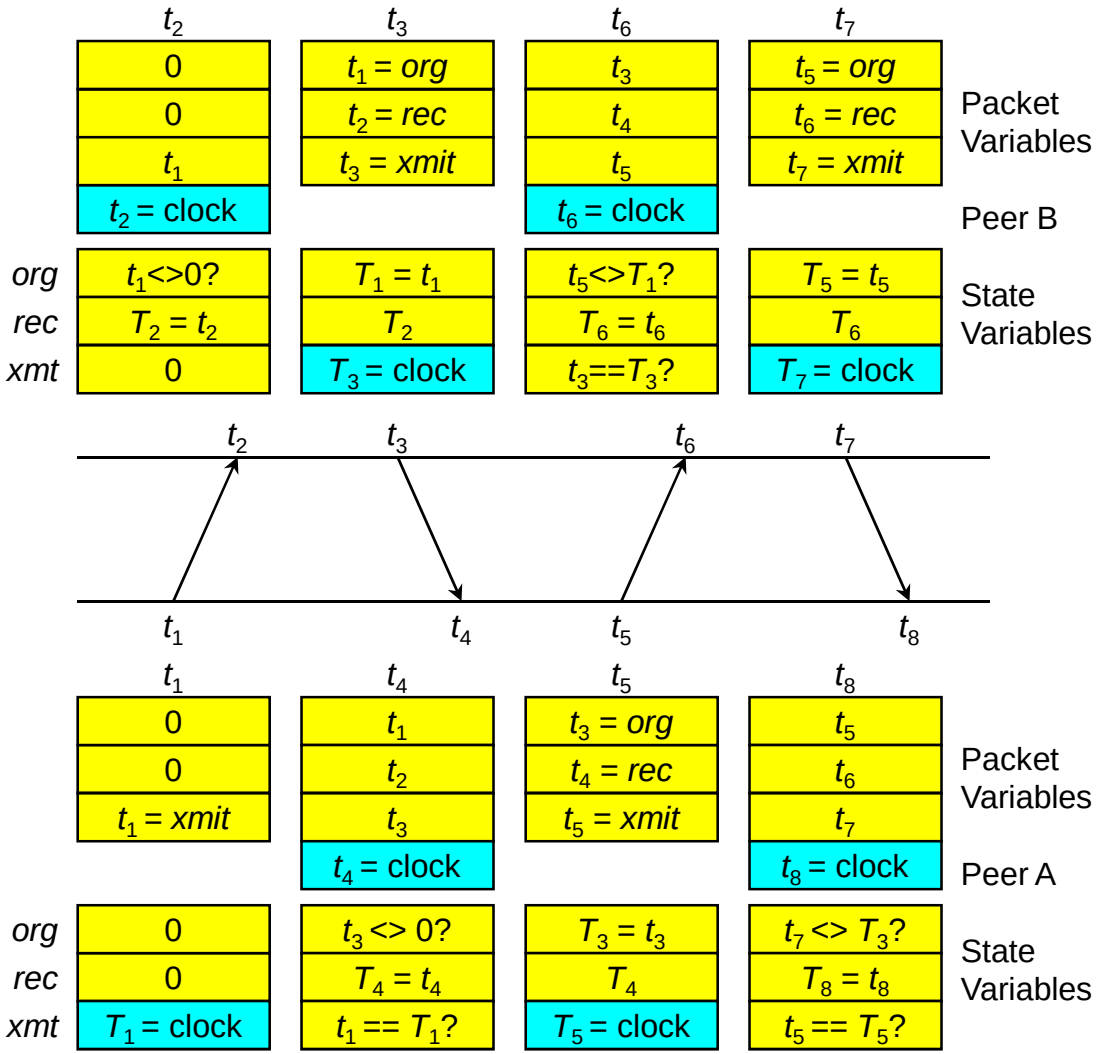


Process decomposition



- Peer process runs when a packet is received.
- Poll process sends packets at intervals determined by the clock discipline process and remote server.
- System process runs when a new peer process update is received.
- Clock discipline process runs at intervals determined by the measured network phase jitter and clock oscillator (VFO) frequency wander.
- Clock adjust process runs at intervals of one second.

NTP one-step on-wire protocol



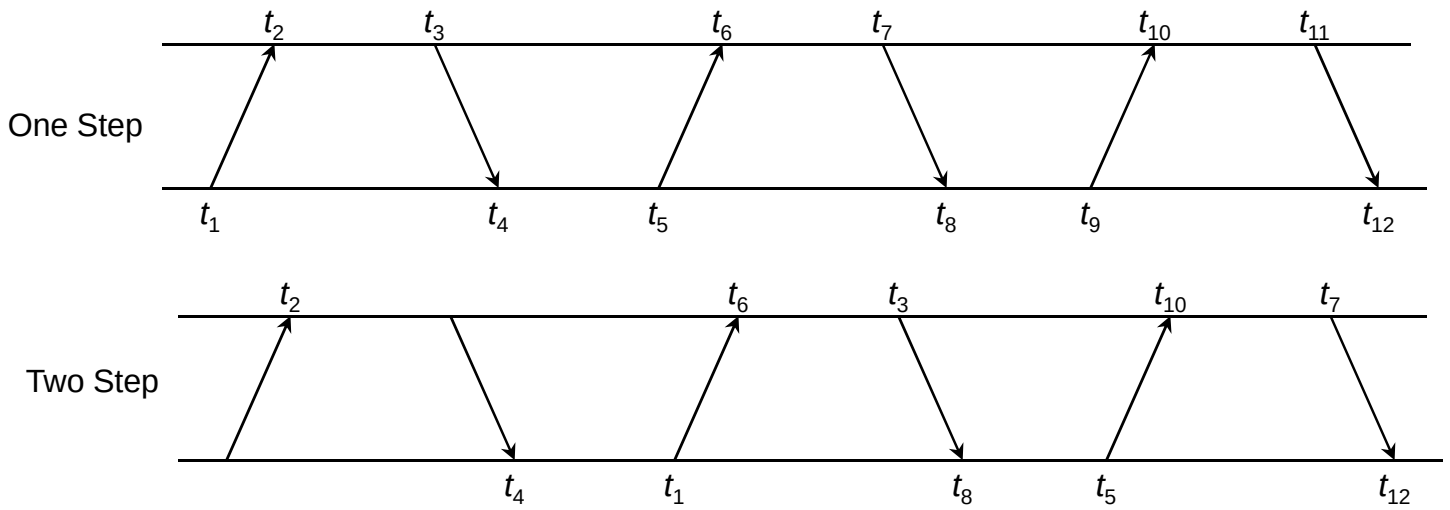
State Variables	
Name	Description
org	originate timestamp
rec	receive timestamp
xmt	transmit timestamp

Packet Header Variables	
Name	Description
t_n	originate timestamp
t_{n+1}	receive timestamp
t_{n+2}	transmit timestamp
t_{n+3}	destination timestamp

$t_7 <> T_3$? *org* Duplicate Test

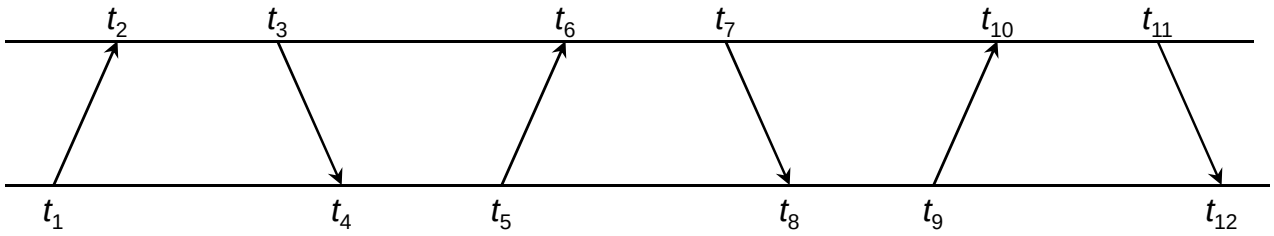
$t_5 == T_5$? *xmt* Bogus Test

Timestamp interleaving



- In the diagrams, transmit timestamps carry odd numbers, receive timestamps carry even numbers.
  - Receive timestamps are available immediately.
  - In one-step mode, transmit timestamps are conveyed in the transmitted packet.
  - In two-step mode, transmit timestamps are conveyed in the following transmitted packet.
  - It takes two roundtrips to accumulate all four timestamps

NTP two-step on-wire protocol



	t_1	t_4	t_5	t_8	t_9	t_{12}	
	0	0	t_3 = *org*	t_5	t_7 = *org*	t_9	Packet Variables
	0	t_2	t_4 = *rec*	t_6	t_8 = *rec*	t_{10}	
	0	0	t_5 = *xmit*	t_7 (T_3)	t_9 = *xmit*	t_{11}(T_7)	
		t_4 = clock		t_8 = clock		t_{12} = clock	
org	0	t_3 <> 0?	$T_3 = t_3$	t_7 <> T_3?	$T_7 = t_7$	t_{11} <> T_7?	State Variables
rec	0	$T_4 = t_4$	T_4	$T_8 = t_8$	T_8	$T_{12} = t_{12}$	
xmt	0	t_1 == T_1?	T_1	t_5 == T_1?	T_5	t_9 == T_5?	
porg	T_1= clock	T_1	T_5 = clock	T_5	T_9 = clock	T_9	Extended State Variables
prec	0	$T_2 = t_2$	T_2	T_6	T_6	T_6	
pxmt	*0*	$T_4 = t_4$	T_4	T_8	T_8	T_8	

Round 1

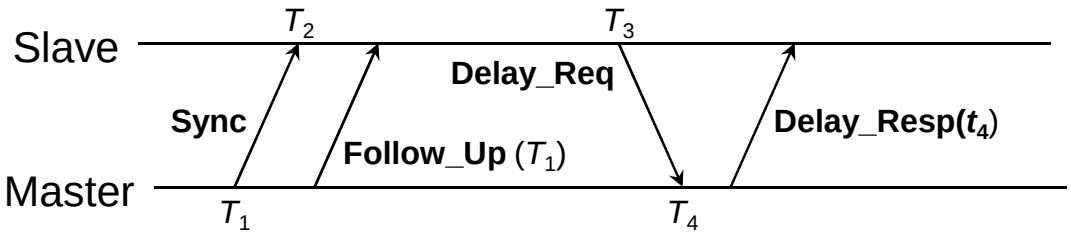
Round 2

Round 3

$$\theta = \frac{1}{2}[(T_2 - T_1) + (T_3 - T_4)]$$
$$\delta = (T_4 - T_1) - (T_3 - T_2)$$

$$\theta = \frac{1}{2}[(T_6 - T_5) + (T_7 - T_8)]$$
$$\delta = (T_8 - T_5) - (T_7 - T_6)$$

IEEW 1588 (PTP) master-slave protocol



- Ethernet NIC hardware strikes a timestamp after the preamble and before the data separately for transmit and receive.
- In each round master sends Sync message at *T*1; slave receives at *T*2.
- In one-step variant *T*1 is inserted just before the data in the Sync message; in two-step variant t1 is sent later in a Follow_Up message.
- Slave sends Delay_Req message at *T*3; master sends Delay_Resp message with *T*4. Compute master offset θ and roundtrip delay δ:

$$\theta = \frac{1}{2}\left[(T_2 - T_1) + (T_3 - T_4)\right] \qquad \delta = (T_4 - T_1) - (T_3 - T_2)$$

Transition matrix

Packet Mode

Association Mode

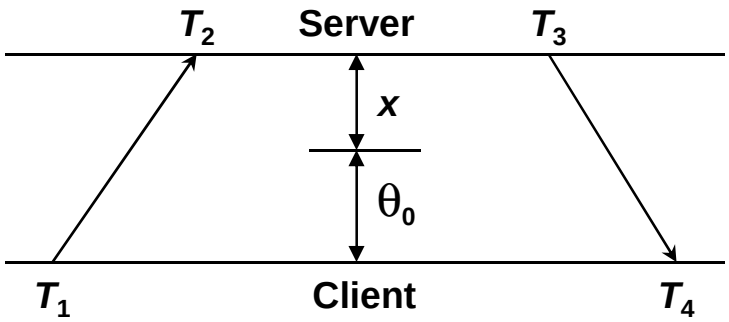
Mode	ACTIVE	PASSIVE	CLIENT	SERVER	BCAST
NO_PEER	NEWPS		FXMIT	NEWMC	NEWBC
ACTIVE	PROC	PROC			
PASSIVE	PROC	ERROR			
CLIENT				PROC	
SERVER					
BCAST					
BCLIENT				ERROR	PROC

The default (empty box) behavior is to discard the packet without comment.

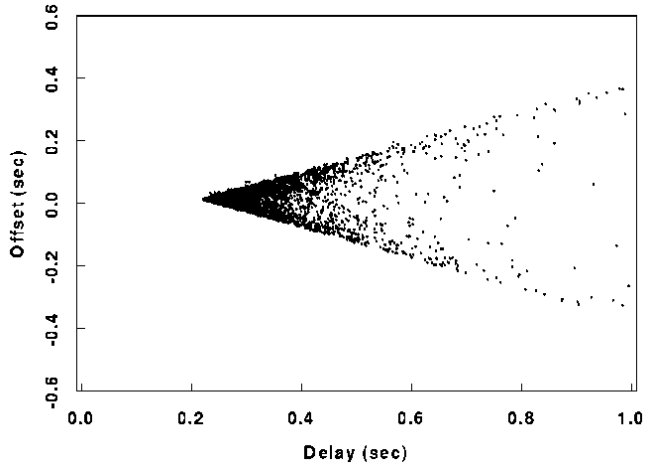
Packet sanity tests

Test	Comment	Code	Condition	Routine
	Packet Flashers			
drop	implementation error	none	$T_3 = 0$ or ($T_1 = 0$ and $T_2 \neq 0$) or ($T_1 \neq 0$ and $T_2 = 0$)	receive
1	duplicate packet	pkt_dupe	$T_3 = xmt$	receive
2	bogus packet	pkt_bogus	$T_1 \neq org$	receive
3	invalid timestamp	pkt_proto	*mode* ≠ BCST and $T_1 = 0$ and $T_2 = 0$	receive
4	access denied	pkt_denied	access restricted, untrusted key, etc.	receive
5	authentication error	pkt_auth	MD5 message hash fails to match message digest	receive
6	peer not synchronized	pkt_unsync	*leap* = 11 or *stratum* >= MAXSTRAT or $T_3 <$ *reftime*	packet
7	invalid distance	pkt_dist	$\Delta_R < 0$ or $E_R < 0$ or $\Delta_R / 2 + E_R >$ MAXDISP	packet
8	autokey keystream error	pkt_autokey	MD5 autokey hash fails to match previous key ID	receive
9	autokey protocol error	pkt_crypto	key mismatch, certificate expired, etc.	receive
	Peer Flashers			
10	peer stratum exceeded	peer_stratum	stratum > sys_stratum in non-symmetric mode	accept
11	peer distance exceeded	peer_dist	distance greater than MAXDIST	accept
12	peer synchronization loop	peer_loop	peer is synchronized to this host	accept
13	peer unfit for synchronization	peer_unfit	unreachable, unsynchronized, noselect	accept

Clock filter algorithm

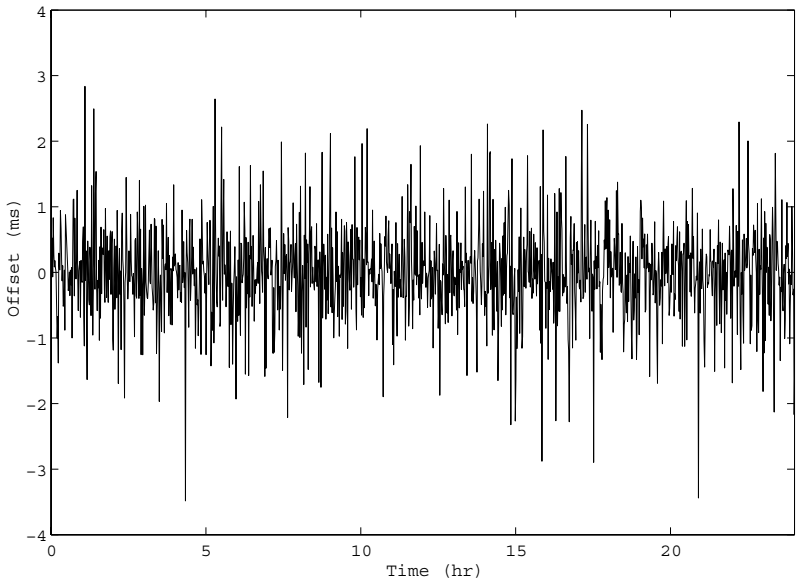


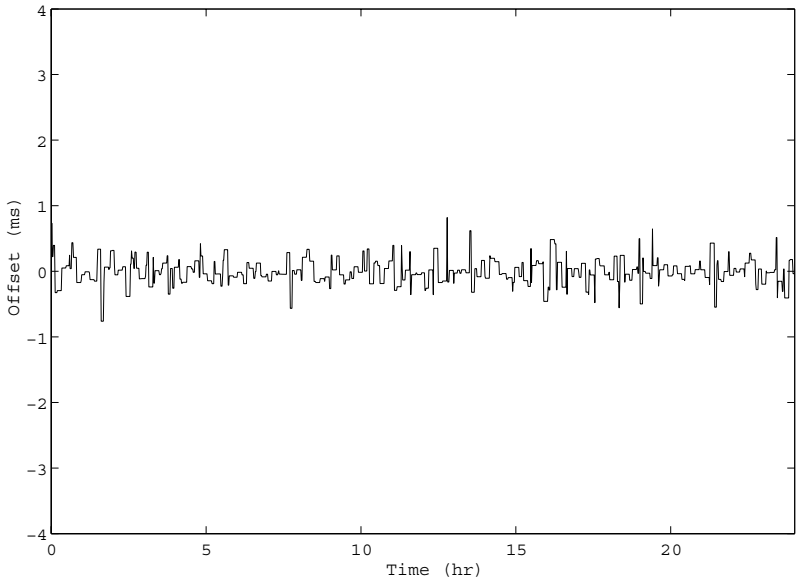
$$\theta = \tfrac{1}{2}[(T_2 - T_1) + (T_3 - T_4)]$$
$$\delta = (T_4 - T_1) - (T_3 - T_2)$$



- The most accurate offset θ_0 is measured at the lowest delay δ_0 (apex of the wedge scattergram).
- The correct time θ must lie within the wedge $\theta_0 \pm (\delta - \delta_0)/2$.
- The δ_0 is estimated as the minimum of the last eight delay measurements and (θ_0, δ_0) becomes the peer update.
- Each peer update can be used only once and must be more recent than the previous update.

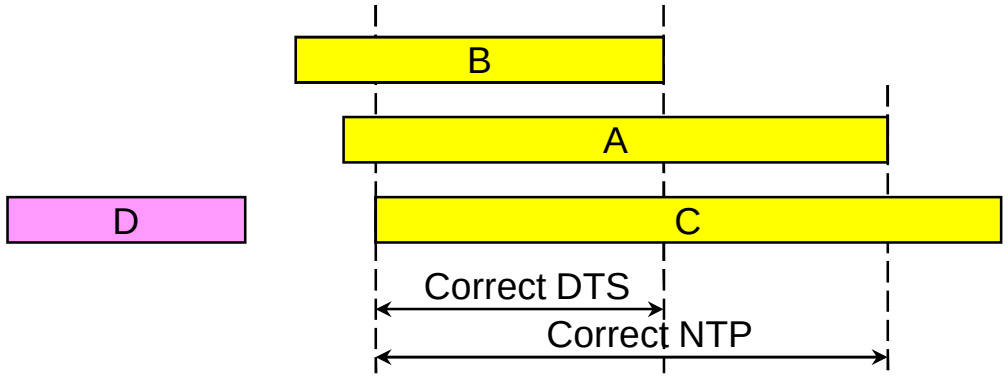
Clock filter performance





- Left figure shows raw time offsets measured for a typical path over a 24-hour period (mean error 724 µs, median error 192 µs)
- Right graph shows filtered time offsets over the same period (mean error 192 µs, median error 112 µs).
- The mean error has been reduced by 11.5 dB; the median error by 18.3 dB. This is impressive performance.

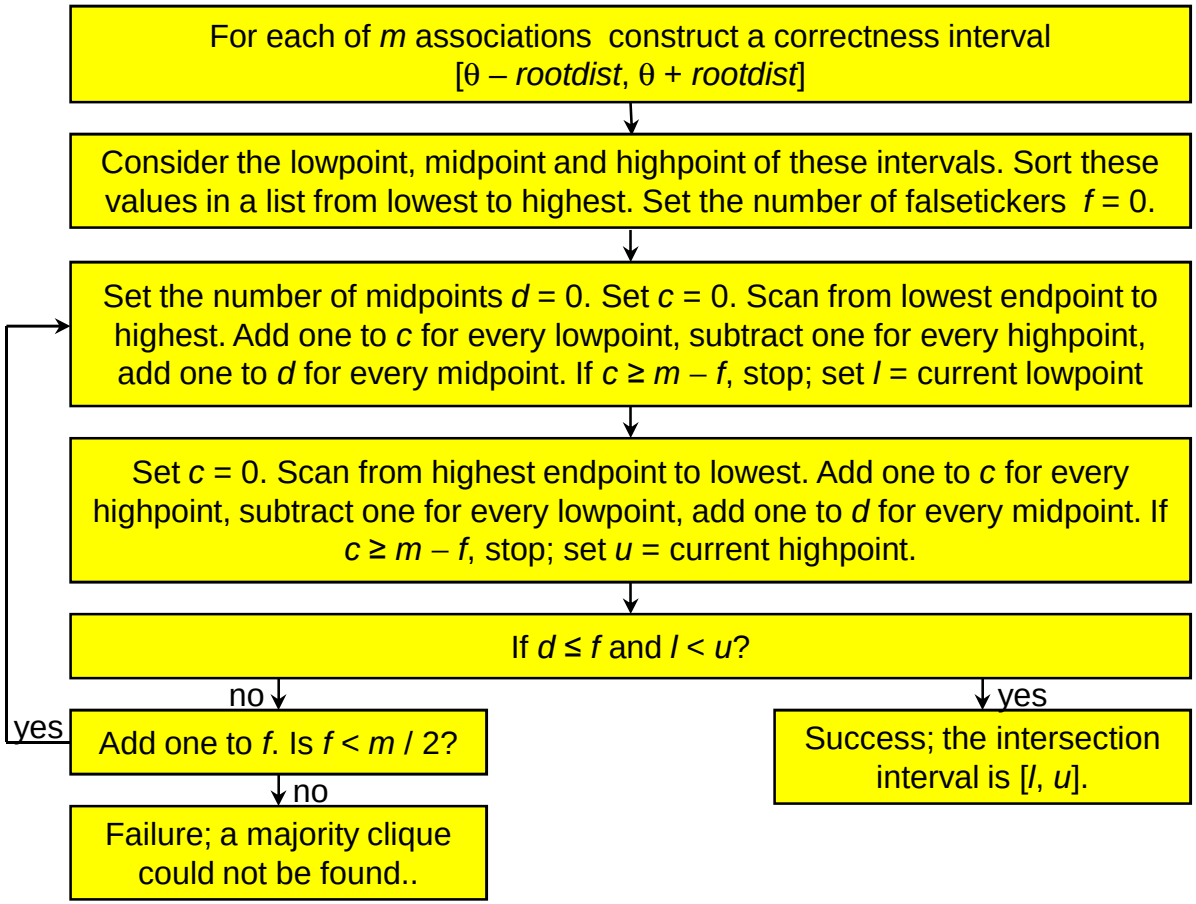
Clock select principles



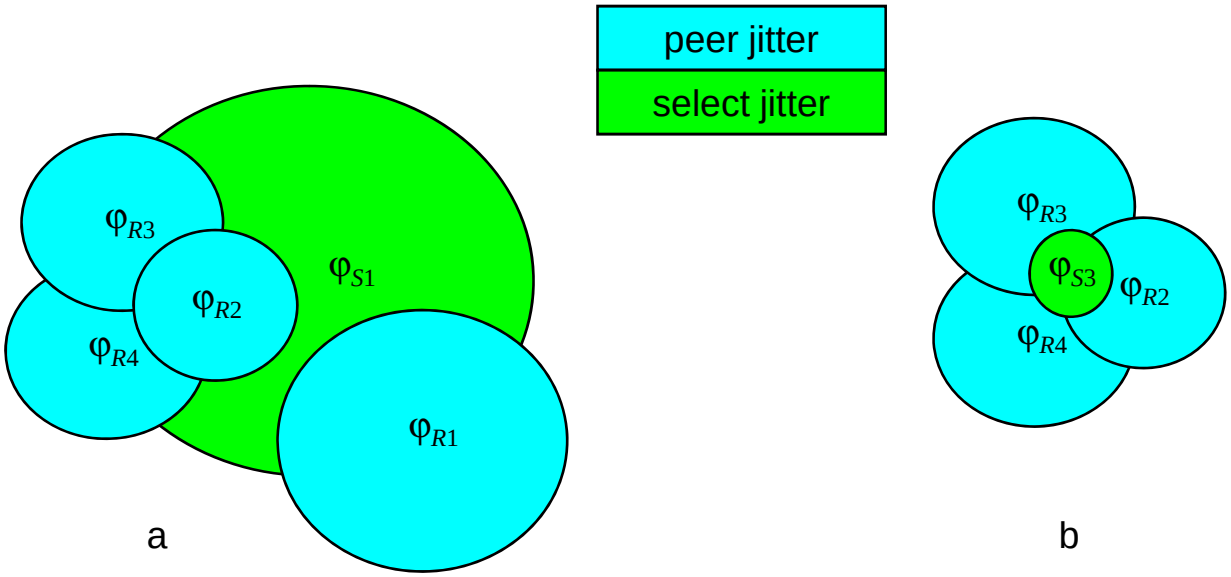
correctness interval = $q - l \pounds q_0 \pounds q + l$
m = number of clocks
f = number of presumed falsetickers
A, B, C are truechimers
D is falseticker

- The correctness interval for any candidate is the set of points in the interval of length twice the synchronization distance centered at the computed offset.
- The DTS interval contains points from the largest number of correctness intervals, i.e., the intersection of correctness intervals.
- The NTP interval includes the DTS interval, but requires that the computed offset for each candidate is contained in the interval.
- Formal correctness assertions require at least half the candidates be in the NTP interval. If not, no candidate can be considered a truechimer.

system process: select algorithm

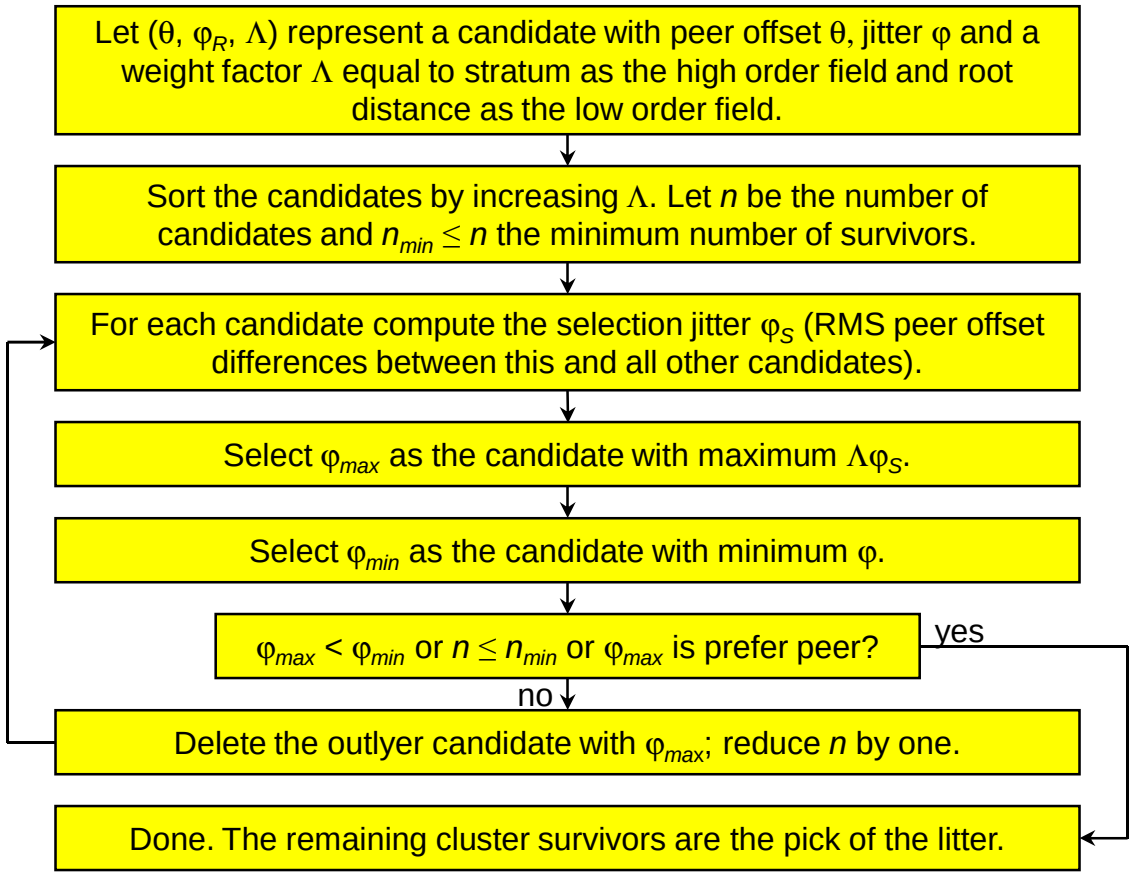


Cluster principles

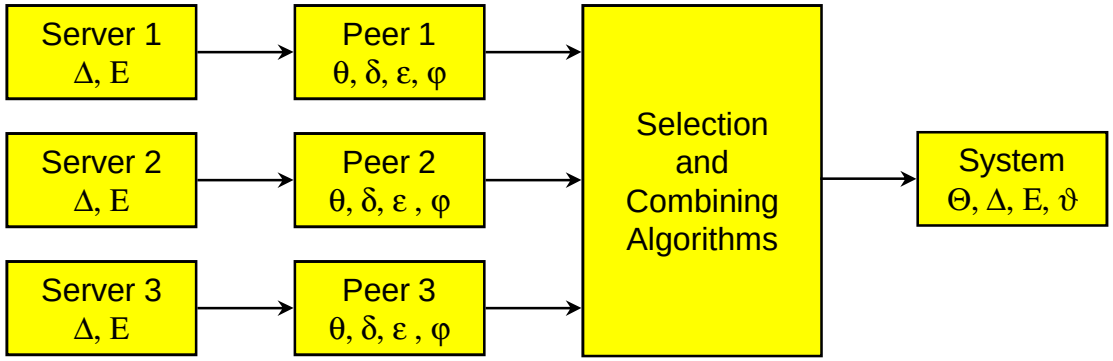


- Candidate 1 is further from the others, so its select jitter φ_{S1} is highest.
- (a) $\varphi_{max} = \varphi_{S1}$ and $\varphi_{min} = \varphi_{R2}$. Since $\varphi_{max} > \varphi_{min}$, the algorithm prunes candidate 1 to reduce select jitter and continues.
- (b) $\varphi_{max} = \varphi_{S3}$ and $\varphi_{min} = \varphi_{R2}$. Since $\varphi_{max} < \varphi_{min}$, pruning additional candidates will not reduce select jitter. So, the algorithm ends with φ_{R2}, φ_{R3} and φ_{R4} as survivors.

system process: cluster algorithm



NTP dataflow analysis



- Each server provides delay Δ and dispersion E relative to the root of the synchronization subtree.
- As each NTP message arrives, the peer process updates peer offset θ, delay δ, dispersion ε and jitter φ.
- At system poll intervals, the clock selection and combining algorithms updates system offset Θ, delay Δ, dispersion E and jitter ϑ.
- Dispersions ε and E increase with time at a rate depending on specified frequency tolerance ϕ.

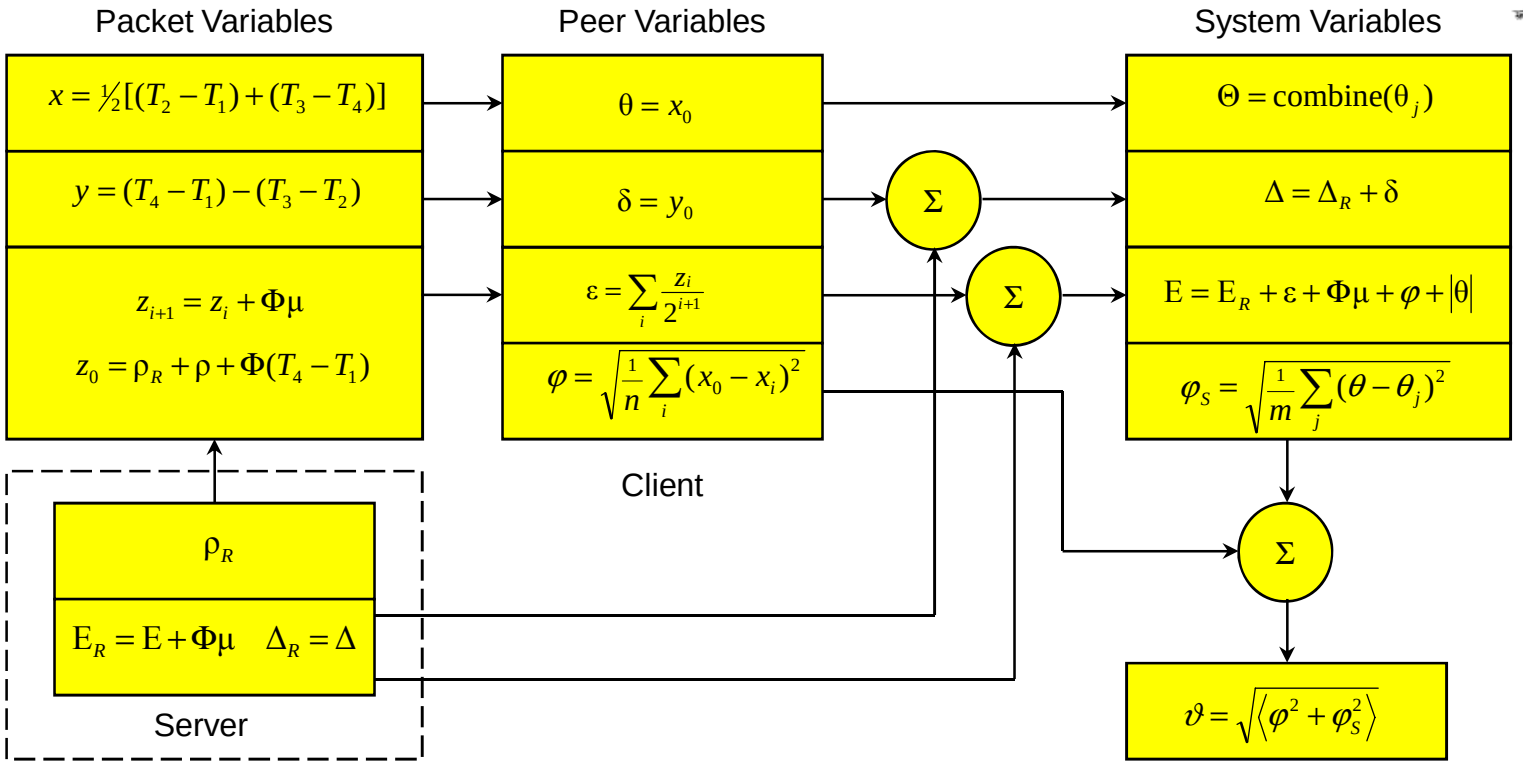
Error budget - notation

- Constants (peers *A* and *B*)
  - ρ maximum reading error
  - ϕ maximum frequency error
  - w dispersion normalize: 0.5
- Packet variables
  - Δ_B peer root delay
  - E_B peer root dispersion
- Sample variables
  - T_1, T_2, T_3, T_4 protocol timestamps
  - x clock offset
  - y roundtrip delay
  - z dispersion
  - τ interval since last update
- System variables
  - Θ clock offset
  - Δ root delay
  - E root dispersion
  - φ_s selection jitter
  - φ jitter
  - τ interval since last update
  - m number of peers
- Peer variables
  - θ clock offset
  - δ roundtrip delay
  - ε dispersion
  - φ_r filter jitter
  - n number of filter stages
  - τ interval since last update

Definitions

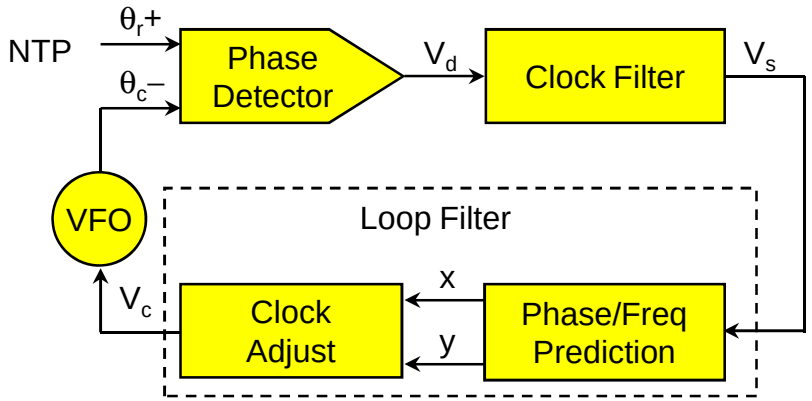
- Precision: elapsed time to read the system clock from userland.
- Resolution: significant bits of the timestamp fraction.
- Maximum error: maximum error due all causes (see error budget).
- Offset: estimated time offset relative to the server time.
- Jitter: exponential average of first-order time differences
- Frequency: estimated frequency offset relative to UTC.
- Wander: exponential average of first-order frequency differences.
- Dispersion: maximum error due oscillator frequency tolerance.
- Root delay: accumulated roundtrip delay via primary server.
- Root dispersion: accumulated total dispersion from primary server.
- Estimated error: RMS accumulation from all causes (see error budget).

Time values and computations



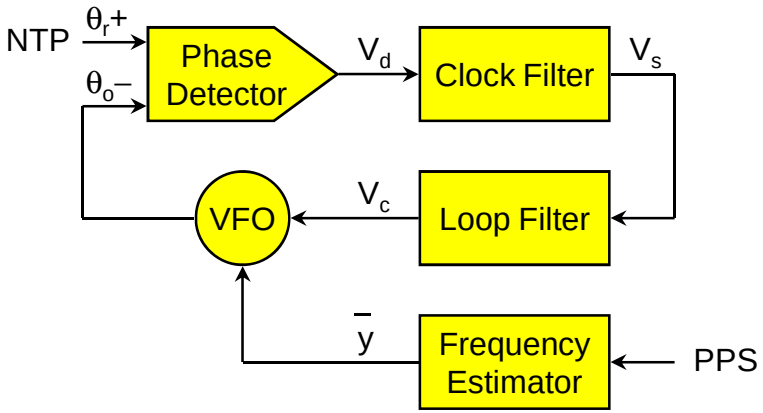
- Packet variables are computed directly from the packet header.
- Peer variables are groomed by the clock filter.
- System variables are groomed from the available peers.

Clock discipline algorithm



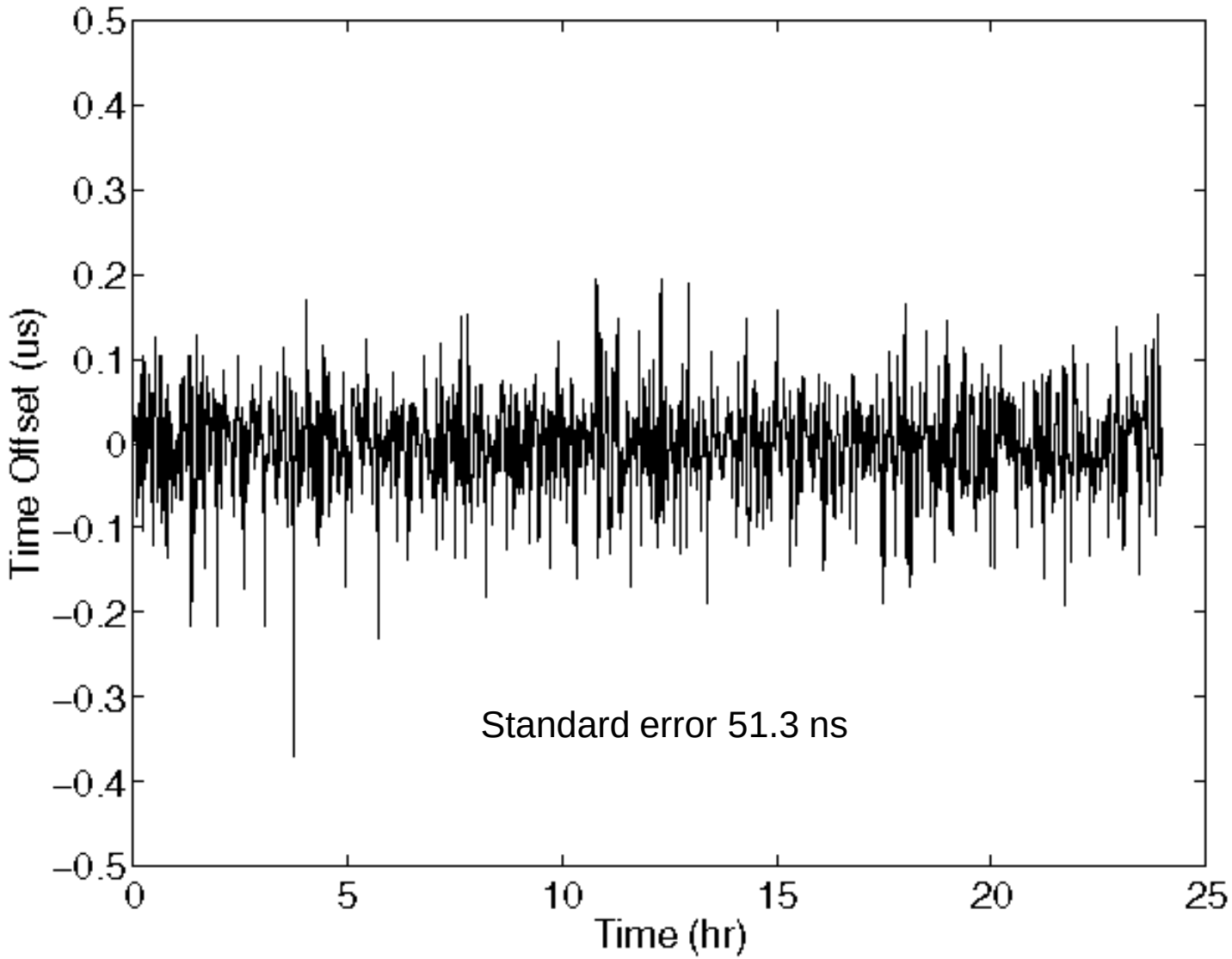
- V_d is a function of the phase difference between NTP and the VFO.
- V_s depends on the stage chosen on the clock filter shift register.
- *x* and *y* are the phase update and frequency update, respectively, computed by the prediction functions.
- Clock adjust process runs once per second to compute V_c, which controls the frequency of the local clock oscillator.
- VFO phase is compared to NTP phase to close the feedback loop.

NTP clock discipline with PPS steering



- NTP daemon disciplines variable frequency oscillator (VFO) phase V_c relative to accurate and reliable network sources.
- Kernel disciplines VFO frequency *y* to pulse-per-second (PPS) signal.
- Clock accuracy continues to be disciplined even if NTP daemon or sources fail.
- In general, the accuracy is only slightly degraded relative to a local reference source.

Measured PPS time error for Alpha 433



Further information

- NTP home page http://www.ntp.org
  - Current NTP Version 3 and 4 software and documentation
  - FAQ and links to other sources and interesting places
- David L. Mills home page http://www.eecis.udel.edu/~mills
  - Papers, reports and memoranda in PostScript and PDF formats
  - Briefings in HTML, PostScript, PowerPoint and PDF formats
  - Collaboration resources hardware, software and documentation
  - Songs, photo galleries and after-dinner speech scripts
- Udel FTP server: ftp://ftp.udel.edu/pub/ntp
  - Current NTP Version software, documentation and support
  - Collaboration resources and junkbox
- Related projects http://www.eecis.udel.edu/~mills/status.htm
  - Current research project descriptions and briefings