Performance optimization of detector electronics for millimeter laser ranging

S. Cova[+], *A. Lacaita*[+] *and G. Ripamonti*[*]

[+]Politecnico di Milano, Dipartimento di Elettronica e Informazione and CEQSE-CNR,
Piazza L. da Vinci 32 - 20133 Milano (Italy)

*Universita' degli Studi di Milano, Dipartimento di Fisica
Via Celoria 16 - 20133 Milano (Italy)

(invited paper)

ABSTRACT

The front-end electronic circuitry plays a fundamental role in determining the performance actually obtained from ultrafast and highly sensitive photodetectors. We deal here with electronic problems met working with Microchannel Plate photomultipliers (MCP-PMTs) and Single Photon Avalanche Diodes (SPADs) for detecting single optical photons and measuring their arrival time with picosecond resolution. The performance of available fast circuits is critically analyzed. Criteria for selecting the most suitable electronics are derived and solutions for exploiting at best the detector performance are presented and discussed

I. INTRODUCTION

Laser ranging applications with millimeter resolution require to measure the time of flight of single photons with precision better than 30ps root mean square (rms), that is, with better than 70ps full-width at half maximum (FWHM) of the resolution curve. Two photodetector types can attain single photon sensitivity and picosecond resolution: proximity-focused Microchannel Plates (MCPs) [1-3] and Single Photon Avalanche Diodes (SPADs) [4-7]. In both cases the front-end electronics associated to the detector plays a fundamental role. In order to take full advantage of the detector timing performance, the electronic pulse processing should be carefully optimized.

In set ups where MCPs are employed, the fast preamplifier and the constant fraction trigger circuit (CFT) are by far the most critical electronic components. In this paper we discuss criteria for optimum selection of the preamplifier, taking into account the noise and bandwidth characteristics. We show that preamplifiers with very large bandwidth (3GHz or more) are not to be employed, since they impair the timing performance [8]. We analyze problems met by constant fraction triggers working with

subnanosecond pulses from MCPs. We discuss how the performance of available CFTs can be improved by simple modifications that provide a better adjustment of the relevant CFT parameters and/or by suitable pre-filtering of the MCP pulses [9,10].

SPADs are avalanche photodiodes that operate biased above the breakdown voltage in the so-called Geiger-mode [4-7,11]. Their operation is fundamentally different from that of photomultiplier tubes (PMTs) and of ordinary avalanche photodiodes (APDs). The device does not have a linear internal gain, that is, it does not amplify linearly the primary photocurrent. It instead exploits the avalanche process to behave in a way similar to that of a trigger circuit, rather than an amplifier. When one or more photons are detected at a given time, a fast-rising current pulse is generated, with standard amplitude and shape, independent of the number of photons. The leading edge of this pulse marks with very high precision, down to 20ps FWHM, the time of arrival of the photon that has triggered the avalanche [5].

We analyze limitations met working with SPADs in the simple biasing arrangements that employ a ballast resistor to quench the avalanche, called Passive Quenching Circuits (PQCs) [6,7,11]. In order to fully exploit the ultimate SPAD timing performance, the device should be operated with an Active Quenching Circuit (AQC) [12,13]. We introduce a new AQC model, capable of driving the detector in remote position, connected by a coaxial cable [13]. Special care has been devoted to the design of the input stage, in order to minimize the circuit noise and thus reduce to less than 3ps rms the internal time jitter of the circuit, making negligible its contribution to the overall time resolution.

II. SELECTION OF THE OPTIMUM PREAMPLIFIER FOR MCPs

Since the gain of the MCP is limited to 5 10^5, a fast preamplifier with gain higher than 10 must be used between the MCP output and the pulse-timing trigger circuit. This gain makes the noise of the following circuits negligible, in comparison to that of the amplifier. In order to avoid reflections and ringing in the pulse shape, the MCP output must be terminated on a 50 Ohm resistor R_s , as outlined in the equivalent circuit of Fig.1.

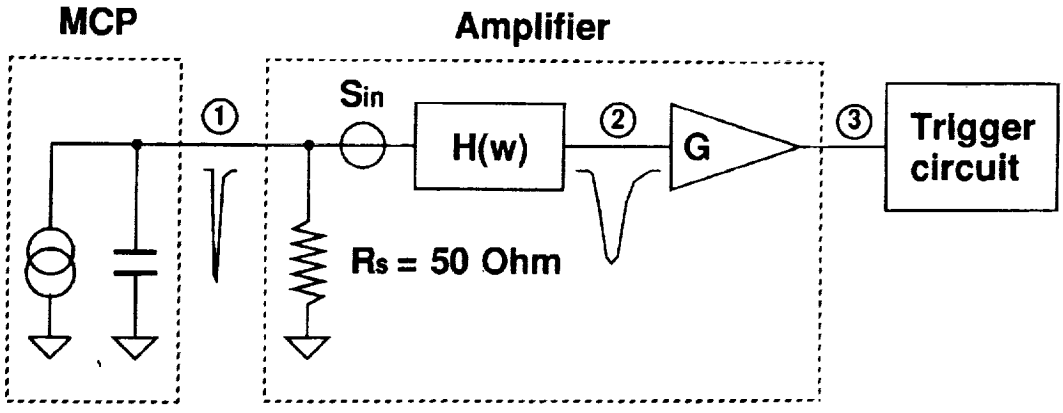

Fig.1 Equivalent circuit for analyzing the effect of the amplifier noise and bandwidth on the pulse time-jitter.

An accurate analysis of the time-jitter contribution arising from the electronic noise has been carried out [8] and we report here the main results. The action of the amplifier on the MCP pulses is described (see Fig.1) by two blocks: a transfer function in the frequency f domain $H_A(f)$, normalized to unity dc gain, followed by a constant gain G. The time-domain impulse response of the amplifier is $h_A(t) = F^{-1}[H_A(f)]$ (where F^{-1} denotes the inverse Fourier transform and $h_A(t)$ is normalized to unit area). S_i is the spectral density of the equivalent input noise generator of the amplifier, assumed to be gaussian. As sketched in Fig.2, the noise causes a random shift of the actual crossing time of the trigger threshold.

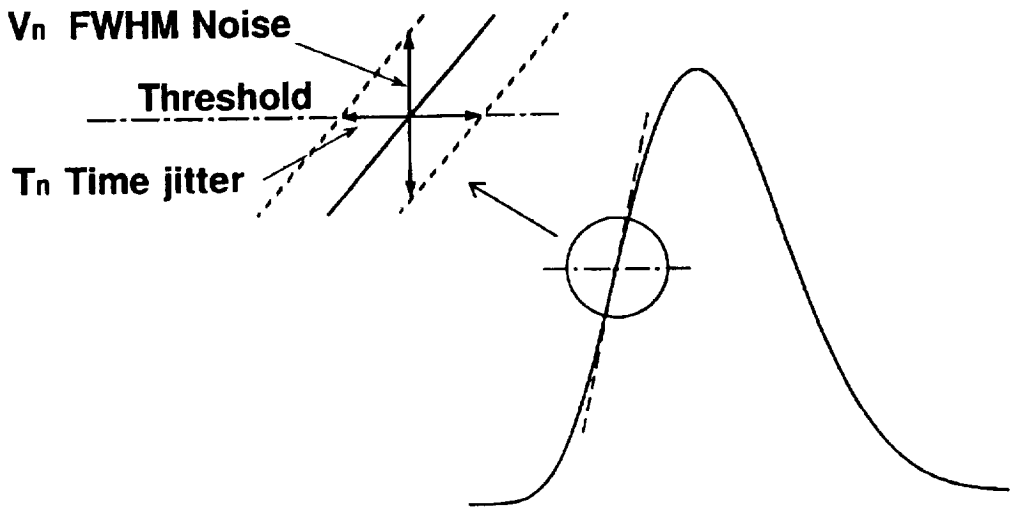

Fig.2 Effect of the electronic noise on the triggering time.

Let us denote by v_n the rms value and by V_n the FWHM of the amplitude distribution of the noise, and by r the slope of the ideal, noise-free pulse at the triggering threshold. The FWHM T_n of the additional time jitter is given by

$$T_n = \frac{V_n}{r} = 2.36\,\frac{v_n}{r} \qquad (1)$$

This time jitter at the comparator threshold (point 3 in Fig.1) is equivalently evaluated at the input of the gain stage G (point 2 in Fig.1).

In first instance, the noise spectrum can be considered white, that is, with constant spectral amplitude $S_i = a^2$. Let us denote by B_n the noise bandwidth, proportional to the signal bandwidth B_A (3dB down bandwidth)

$$B_n = K_n B_A \qquad (2)$$

with constant K_n depending on the shape of the frequency response $H_A(f)$. We have

$$v_n = a\sqrt{B_n} = a\sqrt{K_n B_A} \qquad (3)$$

The slope r also increases with B_A , and it can be easily seen that there is a minimum in the plot of T_n versus B_A. Let us first consider the high bandwidth side, where the B_A values are high enough to have risetime practically equal to that of the MCP pulse. On that side, T_n goes up as $\sqrt{B_A}$, since the slope r is unaffected and the noise v_n increases as $\sqrt{B_A}$. Let us now consider the low bandwidth side, where the pulse-risetime T_r is fully dominated by B_A, namely, T_r is about $1/(3\ B_A)$. On this side, T_n goes down as $B_A^{-3/2}$ when B_A is increased, since the slope r increases as B_A^2 and noise increases as $\sqrt{B_A}$. A minimum of T_n will therefore be found at an intermediate value of B_A , at which the relative rate of increase of the pulse slope r will be equal to that of the noise V_n. More accurate quantitative results can be obtained by considering the actual voltage waveform. Let us denote by $V_D(t)$ the voltage pulse at the 50 Ohm output of the MCP detector

$$V_D(t) = Q\ h_D(t) \tag{4}$$

where Q denotes the area of $V_D(t)$ and $h_D(t)$ is normalized to unit area. The actual voltage waveform V(t) at the input of the gain stage G (point 2 in Fig.1), results from the convolution product of the detector pulse $V_D(t)$ and of the amplifier impulse response $h_A(t)$

$$V(t) = Q\ h_D(t) * h_A(t) \tag{5}$$

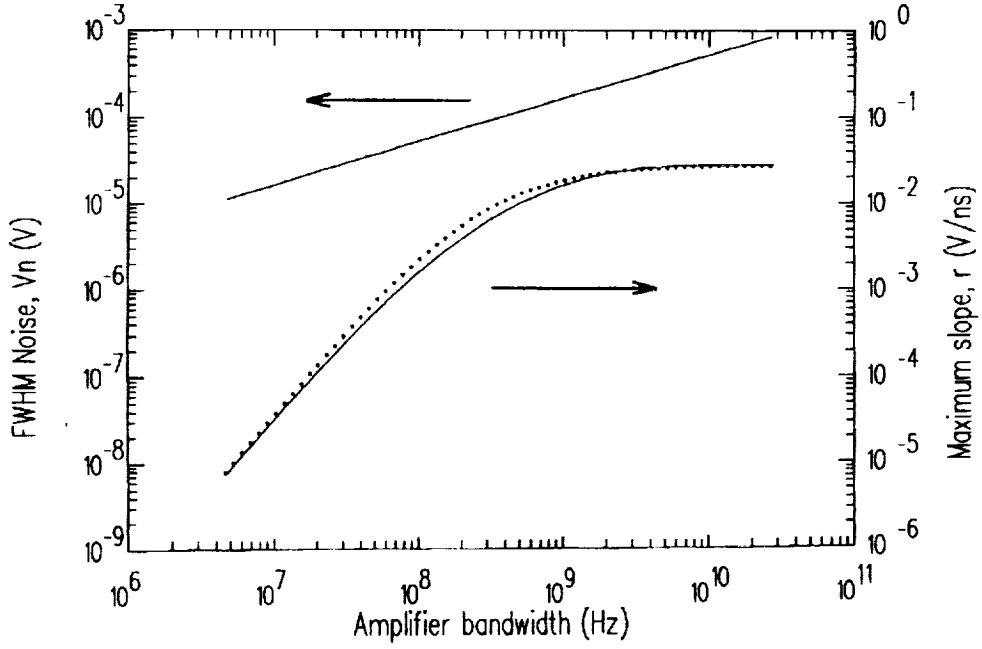

Fig.3 Maximum slope, r, of the amplified pulse and FWHM Vn of the noise amplitude distribution at point 2 in Fig.1 vs amplifier bandwidth B_A. Pulses of a 12μm channel MCP (Hamamatsu R1564U); amplifier having frequency response with two real poles, white input noise generator with rms density $a = 2nV\ Hz^{-1/2}$. Results of detailed computations of the pulse waveform (full line) are compared with the approximation discussed in Ref. [8] (dotted line).

The shape of V(t) and the corresponding maximum slope r can be obtained by numerical computations, accurately taking into account the characteristics of the detector pulse and of the amplifier impulse response [8]. A simple approximation can also be employed to obtain a sufficiently good estimate of r. Essentially, it consists in computing the slope r as result of a weighted average of the corresponding slopes of the detector pulse and of the amplifier impulse response [8]. Fig.3 reports the computed behavior of r and V_n versus amplifier bandwidth for a typical case. Fig.4 illustrates for another case the detailed behavior of the time jitter versus amplifier bandwidth. A broad minimum is found in all cases considered, centered at an optimum bandwidth value B_{Aopt} , which in all cases is well below 2GHz and mostly is around 1GHz .

The analysis carried out with a white noise spectrum leads to definite conclusions. Selecting the amplifier for very high bandwidth and paying minor attention to the noise is not only useless, but even disadvantageous. The noise has primary importance, since the time jitter is proportional to the root-mean square spectral density a. The optimum bandwidth value B_{Aopt} is markedly lower than that suggested by the criterion of keeping the risetime of the amplified pulse very near to the original risetime of the detector pulse. Even for the fastest available MCP detector, B_{Aopt} does not exceed 1.7GHz. Satisfactory results can be obtained with bandwidth values remarkably lower than B_{Aopt}, even by a factor of two. On the basis of these results, it was concluded that amplifiers employing fast bipolar transistors (BJTs) provide better performance than amplifiers based on ultrafast metal-semiconductor field-effect transistors (MESFETs). In fact, the higher bandwidth of MESFETs is not required and, working with a 50 Ohm source, the lower input current noise is not important. Furthermore, MESFETs may have higher components in the low-frequency noise spectrum. However, it must be taken into account that the input noise spectrum of BJTs also contains a high frequency component proportional to f^2. A detailed analysis [8] shows that the f^2 noise component causes a steeper increase of V_n for rising B_A, thereby shifting to lower B_A value the minimum of T_n and making much steeper the rise of T_n on the high B_A side. Fig.5 illustrates a typical case. It is worth stressing that the effect is more marked and the shift of the minimum is greater for transistors having lower white noise component (that is, lower value of a).

These results further support and enhance the conclusions drawn in the white noise analysis. The presence of a f^2 component in the noise spectrum of BJTs i) shifts to even lower value the optimum amplifier bandwidth B_{Aopt} ; ii) makes remarkably more severe the penalty for working with amplifier bandwidths higher than the optimum one ; iii) has almost negligible influence on the results obtained with amplifier bandwidth lower than B_{Aopt}.

In conclusion, the quantitative analysis demonstrates that, by using available low-noise high-frequency bipolar transistors in rationally selected operating conditions, the additional time jitter due to the circuit noise can be kept below 5 ps FWHM. In any case of interest, the behavior of the jitter versus the amplifier bandwidth can be fairly simply analyzed by using the approximate approach introduced in Ref.8. This approach just requires the knowledge of a few numerical parameters characterizing the pulse shapes involved, which can be obtained by analytical representation or by measurements of the waveforms.

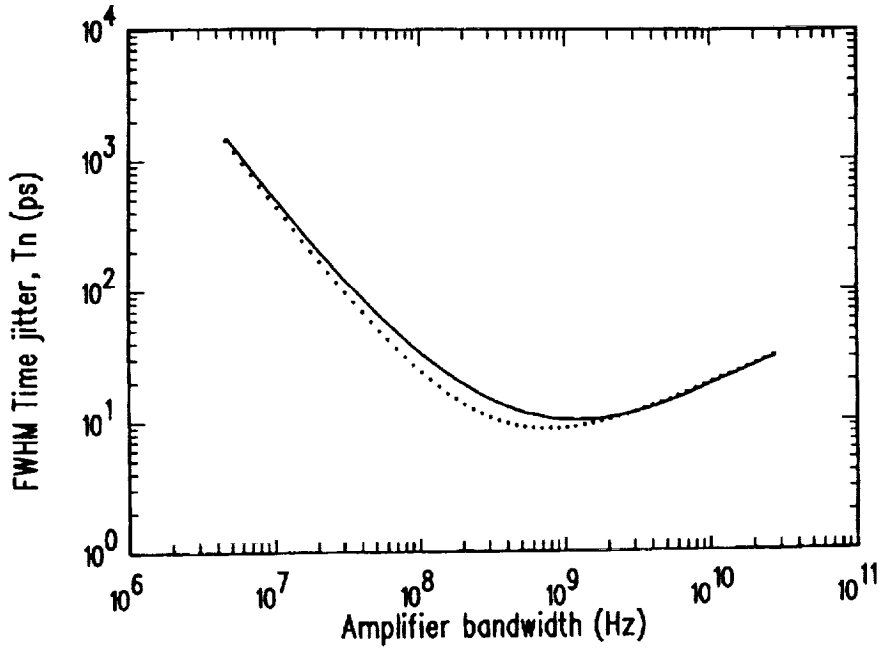

Fig.4 Additional FWHM time jitter T_n due to the noise vs amplifier bandwidth B_A, for the case of Fig.3. Results of detailed computations of the pulse waveform (full line) are compared to the approximation discussed in Ref. [8] (dotted line).

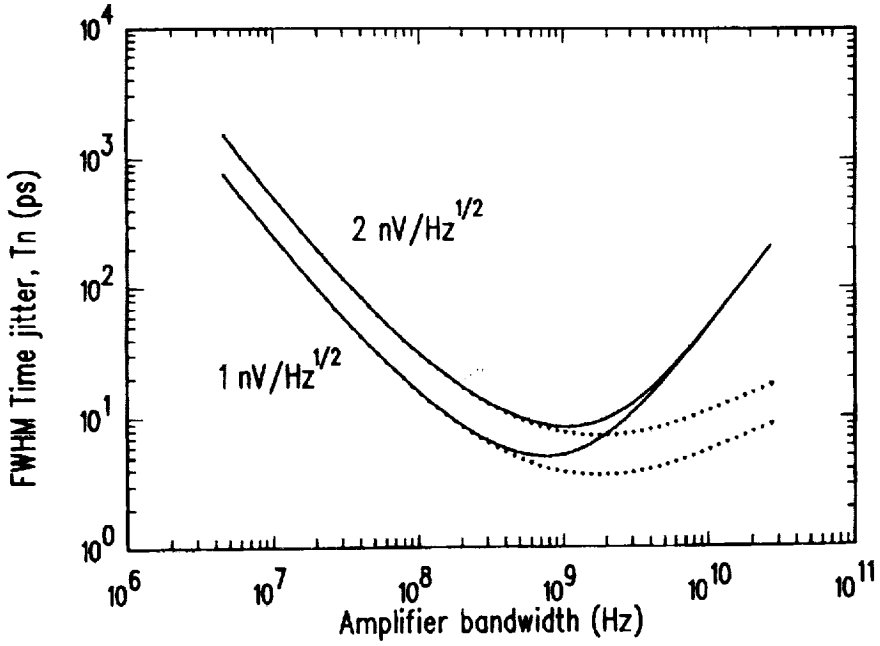

Fig.5 Effect of the f^2 spectral noise component in bipolar transistors (BJT). Pulses of a 6μm-channel MCP (Hamamatsu R2809U). Amplifier with two real poles in the frequency response, input BJT having f_T=5GHz (see Ref.8). Results with the total spectrum (full lines) are compared with those computed with only the white noise term (dotted lines) for two cases: minimum noise a=1nV $Hz^{-1/2}$ and moderately low noise a=2nV $Hz^{-1/2}$

III. CONSTANT FRACTION TRIGGERS WITH MCPs

Since single-photon pulses of photomultiplier tubes (PMTs) have statistically fluctuating amplitude, constant-fraction trigger circuits (CFT) are normally employed for accurate timing [14]. However, with the subnanosecond signals of MCPs non-ideal CFT behavior is observed. A residual amplitude-dependent time-walk sets the ultimate resolution in photon timing. A quantitative analysis of the problem has been carried out and will be here summarized [10].

Fig. 6 illustrates the basic structure and the principle of operation of a CFT circuit. The triggering point of the fast comparator is determined by the crossing of two replicas of the amplified MCP pulse, the first attenuated by a factor k, the second one delayed by a time T_D.

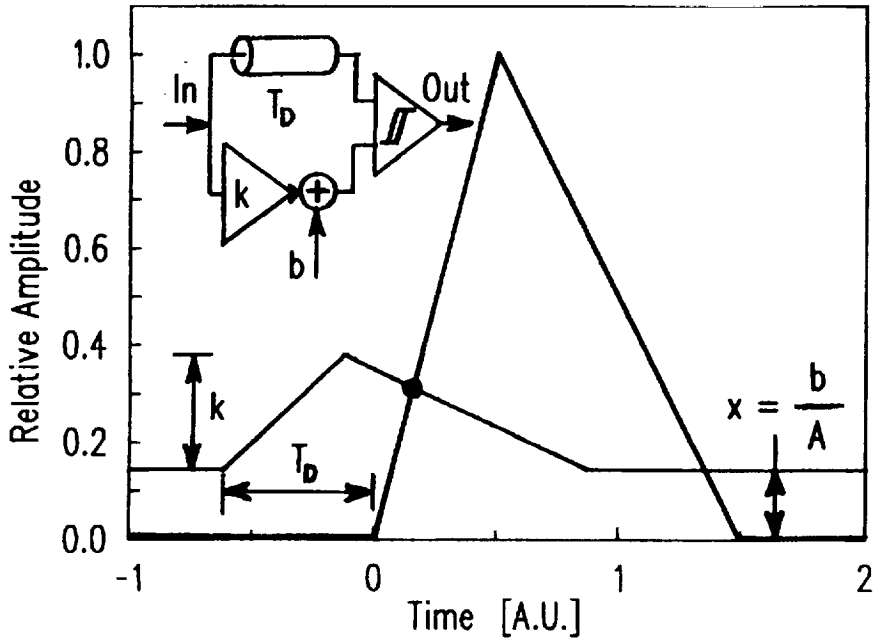

Fig.6 Working principle of Constant Faction Triggers (CFTs) based on a fast comparator with differential input. Block diagram (inset) and pulse waveforms at the comparator inputs (normalized to unit peak amplitude) are depicted in simplified form.

Crossing occurs when the delayed pulse reaches a given percentage of the peak. The time-walk effect is eliminated, since the triggering time is independent of the actual pulse amplitude A. This is strictly true only if the two waveforms have the same baseline level. However, setting both inputs at the same bias voltage level is unacceptable, since it causes the comparator to oscillate. A small dc voltage offset b is necessary; in practice, about 10 mV with modern fast comparators. This constant offset b causes a deviation from the ideal CFT operation, since it causes the crossing time to walk as A is varied. The slower is the rise of the delayed waveform, the larger is the walk. The problem is quantitatively analyzed making reference to the pulse waveform g(t) normalized at unit peak amplitude and to the correspondingly normalized baseline offset x. The pulse-amplitude distribution, with probability density p(A), is transformed in a distribution of

x, with probability density q(x).

$$x = \frac{b}{A} \tag{6}$$

$$q(x) = p(A) \left| \frac{dA}{dx} \right| = p(A) \frac{A^2}{b} \tag{7}$$

The lower threshold level A_L set for accepting the MCP pulses is translated in an upper limit x_H; conversely, the accepted maximum pulse height A_H sets a lower limit x_L. The relation between x and the time walk is readily derived. The crossing time t_c along the waveform g(t) is defined by:

$$x + k\, g(T_D + t_c) = g(t_c) \tag{8}$$

By differentiating with respect to x, we obtain:

$$dx + k \left(\frac{dg}{dt} \right)_{t_c + T_D} dt_c = \left(\frac{dg}{dt} \right)_{t_c} dt_c \tag{9}$$

Denoting by $g'_r = (dg/dt)_{t_c}$ the rising slope and by $g'_f = (dg/dt)_{t_c + T_D}$ the falling slope at the crossing point, we define the intercrossing slope

$$g'_i = g'_r - k\, g'_f \tag{10}$$

and obtain

$$dt_c = \left(\frac{dx}{g'_i} \right) \tag{11}$$

Let us denote by t_o the crossing time for x=0, which corresponds to the ideal CFT case (and is well approximated by the real CFT for pulse amplitude A much higher than the offset b). As A is decreased, the x value is increased and the crossing is shifted from t_o to a later time t_c ; the time walk t_s is $t_s = t_c - t_o$. The equation relating the time walk to x is simply obtained by integrating eq.11. The distribution of x is transformed in a distribution of t_s , with probability density $w(t_s)$. Since $w(t_s)\, dt_s = q(x)dx$ and $dt_s = dt_c$, we obtain from eq.11:

$$w(t_s) = q(x)\, g'_i \tag{12}$$

The actually observed time resolution curve $r_m(t)$ will be the convolution product of this distribution $w(t_s)$ and of the intrinsic resolution curve of the apparatus $r_i(t)$, due to other causes of time dispersion

$$r_m(t) = r_i(t) * w(t) \tag{13}$$

Since w(t) results from an inverse transformation of p(A), it is strongly asymmetric, affected by a long tail towards high t_s values. Its effect in widening the FWHM of $r_m(t)$ is therefore greater than that of a gaussian function having equal FWHM. Taking a linear pulse approximation (as outlined in Fig.6), that is, assuming constant intercrossing slope g'_i, the time walk is proportional to x and the distribution $w(t_s)$ is obtained from q(x) simply by a change of scale. It extends from a lower limit $t_{sL} = g'_i\, x_L = g'_i\, b/A_H$ to an upper limit $t_{sH} = g'_i\, x_H = g'_i\, b/A_L$. With the intercrossing slope g'_i obtained employing the suitable fast preamplifiers (risetime around 400ps, see Sec.II), a fairly small time walk effect would be estimated. For MCP types having 40ps intrinsic FWHM resolution, it would be practically negligible, since the computed FWHM of $r_m(t)$ is less than 45ps. For faster types, having 20ps intrinsic FWHM resolution, it would be moderate, since the computed FWHM of $r_m(t)$ ranges from 25 to 30ps. As a matter of fact, however, in the set-ups actually employed by the experimenters the situation significantly deviates from the linear pulse approximation. As illustrated by Fig.7, this is due to the short duration of the MCP pulses, to their shape and to the CFT setting, in particular to the minimum available value of the delay T_D.

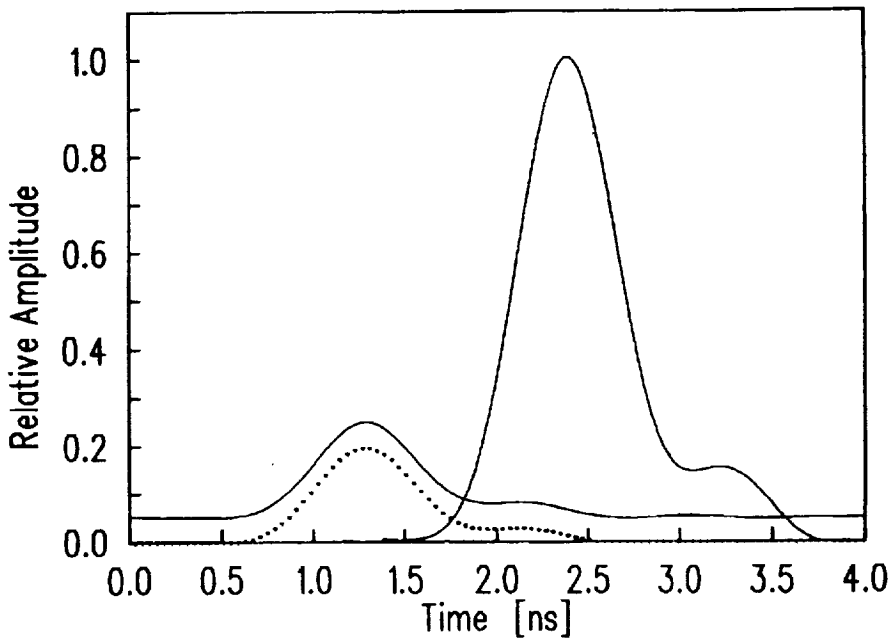

Fig.7 Pulse waveforms at the comparator inputs in the CFT, coming from a 12μm channel MCP (Hamamatsu R1564U), amplified by a HP 8447F fast amplifier. The attenuation is k=0.2, as usual in available CFTs. The delay is T_D=1.1ns, corresponding to minimal external cable length. Actual CFT operation is illustrated by the attenuated pulse with baseline offset (full line); the ideal CFT case is represented by dotted-line waveform.

The intercrossing slope g'_i is a function of the crossing time t_C and cannot be approximated by a constant in eq.s 11 and 12. The transformation from q(x) to $w(t_s)$ is no more linear and a significantly wider $w(t_s)$ results. The situation is particularly bad

when the crossing occurs at the onset of the delayed pulse leading edge, where the rise is slower and the intercrossing slope is correspondingly low. Such a situation can be avoided by making the delay T_D shorter than the duration of the input pulse. In commercial CFT circuits, T_D is mainly determined by the propagation time in a coaxial cable, externally connected to the circuit module and selected by the user. The least obtainable value of T_D is about 1.1ns or slightly less [15], determined by the connectors, the circuit layout and the minimal length possible for the external cable. This was adequate for the pulses of ordinary PMTs, lasting a few nanoseconds, but it is no more sufficient for the subnanosecond pulses of ultrafast MCPs. As illustrated in Fig.7, with T_D=1.1ns the crossing occurs just in the low slope zone of the rising waveform. The operation only roughly approximates the ideal CFT; in fact, it is intermediate between a CFT and a leading edge trigger with small, constant threshold.

Such a situation can be avoided by a suitable selection of parameters in the experiment. The analysis points out the basic criteria for minimizing the time walk effect:

i) the preamplifier gain should be high enough to keep low the value of the upper limit x_H in all cases, even when a small value of the lower threshold A_L is selected for accepting almost all pulses, e.g. 90%of the amplitude distribution p(A).

ii) the delay T_D and attenuation k should be selected for maximizing the intercrossing slope.

The latter criterion is usually not satisfied by industrially produced CFT models. Beside having too long minimum delay, they are normally set to low constant fraction values, around 0.2, which are optimal for timing signals from scintillation detectors of ionizing radiations, but not for timing single photons [10,14]. This is due to historical reasons, since CFTs were originally developed for working with ionizing radiation detectors

A detailed quantitative analysis of the CFT time-walk effect in the conditions of actual experiments can be carried out by means of computer simulation, taking accurately into account the actual shape of the signal waveform processed by the amplifier. It is worth noting that the result of the computation of the crossing time t_c versus normalized baseline offset x can be easily foreseen, by linearly shifting upwards the attenuated waveform in figures like Fig.7 and directly observing the walk of the crossing point. In the following, in order to set in evidence the effects on the time resolution, all the time distribution curves are drawn aligned at the peak value. All FWHM values reported are measured on the complete computed curve $r_m(t)$. In fact, since the shape of the distribution $w(t_s)$ is asymmetrical and far from gaussian, the FWHM value of $r_m(t)$ would be remarkably underestimated by a quadratic composition of the FWHM values of $w(t_s)$ and $r_i(t)$.

Previously published experimental results were analyzed and improvements obtainable by modifying the CFT parameters and/or the pulse shape were evaluated [10]. We first considered cases where a 12micron channel MCP Hamamatsu R1564U is employed with a fast amplifier model HP8447F. The amplitude distribution and signal waveform are reported from Yamazaki et al. [1] and from the manufacturer data sheets

and technical notes [16]. On the basis of published results [1-3] the intrinsic time distribution $r_i(t)$, due to the detector is assumed to be gaussian with 40ps FWHM.

We studied the effect of a filtering stage with a 1ns integrating time constant, interposed between fast amplifier and CFT as proposed and experimented in our laboratory [9]. The wider pulse obtained by filtering makes possible to obtain a higher intercrossing slope with the minimum delay of 1.3ns available in our set up. The computer analysis gives results in perfect agreement with the experiment, confirming that the improvement of the FWHM from 75ps to 55ps was obtained thanks to the reduction of the time walk effect.

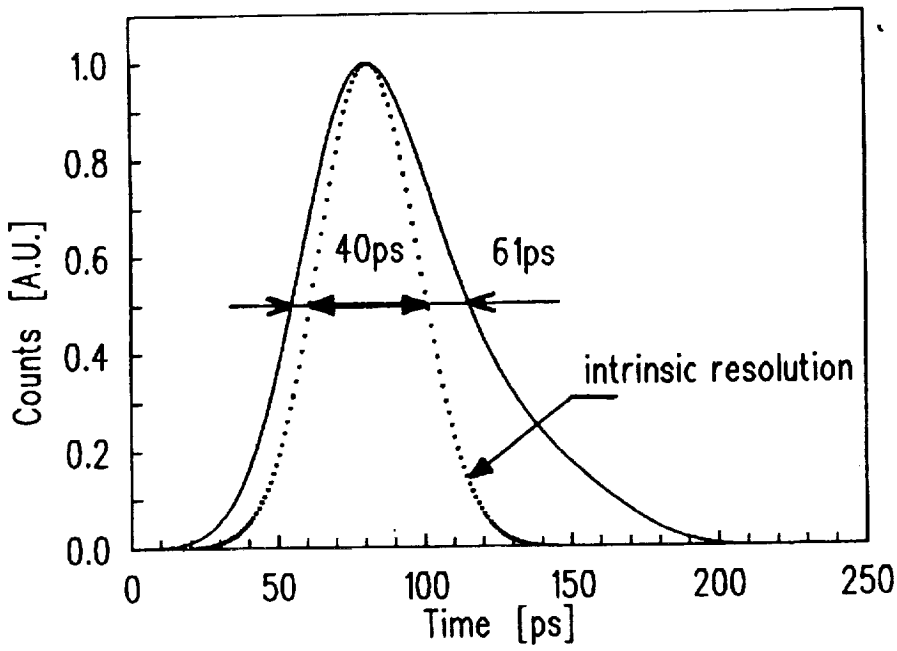

Fig.8 Computed time resolution curve for measurement with low selecting threshold level (accepting almost all pulses), compared to the intrinsic time resolution curve (dotted curve) of a 12 μm channel MCP (Hamamatsu R1564U). The computed FWHM is 61ps; the experimental value in Ref.1 is 63ps.

We analyzed then the experiments reported by Yamazaki et al. [1], where the resolution was measured with two different levels of the auxiliary lower threshold A_L of the CFT, which selects pulses accepted for the time measurement. With low threshold, accepting almost all pulses, they measured 63ps FWHM; the computed value for T_D=1.1ns and k=0.2 was 61ps, as illustrated in Fig.8. The computation also confirmed that with high threshold, accepting only 15% of the pulses, the FWHM suffers negligible degradation with respect to the intrinsic value. For the experiments with low threshold level, we also evaluated the improvement obtainable by reducing the pulse delay T_D and/or the attenuation ratio k, in order to increase the intercrossing slope. It is not difficult to reduce the delay T_D in existing CFTs, without needing to redesign and fabricate new CFT models. The available models can be modified by cutting lines in the printed circuit board and making new internal connections for a shorter delay path, that avoids the external delay cable. By modifying the resistor network that attenuates the prompt waveform (see Fig.6), one can also change the k to higher values. In order to attain k=1, however, major modifications or complete redesigning of the CFT circuit

may be necessary. The results obtained by changing the delay to T_D=0.6ns and/or the attenuation range to k=1 are summarized in Table 1.

We can conclude that with the 12 micron-channel MCP Hamamatsu R1564U it is fairly simple to reduce the time-walk effect to a tolerable or even negligible contribution.

Table 1 FWHM time resolution values obtained in the computer simulation of measurements with Hamamatsu R1564U (intrinsic FWHM resolution 40ps) and CFT circuit with low selecting threshold level (see Fig.8) and different values of the delay T_D and attenuation k.

k	0.2	0.2	1	1
T_D	1.1ns	0.6ns	1.1ns	0.6ns
FWHM	61ps	49ps	50ps	42ps

It is not strictly necessary to design new CFT models, since employing simple auxiliary circuits or making fairly simple modifications to available CFT circuits can be sufficient to the purpose. By employing a simple filter-amplifier to increase the width of the pulse fed to the CFT, without modifying the CFT circuit, the resolution widening can be reduced to less than 40%. By modifying the CFT to reduce the delay, the widening is limited to 22%. If, further to reducing the delay, the attenuation is eliminated (k=1), the widening drops to 5%. It is interesting to note that some older ZCT circuits may therefore be more suitable to single-photon timing than modern CFTs.

We may also note that the results of this analysis suggest how to select a suitable fast amplifier among a set of available models with a given value of the gain-bandwidth product, with criteria in agreement with the conclusions drawn from the analysis of the contribution of the electronic circuit noise to the time jitter. It is clearly better to select a model with moderate bandwidth (1GHz or less) and higher gain, rather than faster models with lower gain. The task of avoiding significant time-walk effect in the resolution becomes tougher with the latest and fastest MCP detectors, having 20ps or better FWHM intrinsic time resolution, such as the 6micron channel MCP model Hamamatsu R2809U. Data for the analysis of this case were obtained from Ref.3 and from the manufacturer data sheet and technical notes [16]. The intrinsic resolution curve $r_i(t)$ is assumed to be gaussian with 20ps FWHM. Since model HP8447F amplifier is very well suited also to this case [8] from the standpoint of the time jitter due to circuit noise, we analyzed set-ups employing such a fast amplifier. Fig.9 illustrates the effect of employing a shorter delay T_D in the CFT. Table 2 summarizes the improvements obtainable with modifications to the CFT circuit.

Table 2. FWHM time resolution values obtained in the computer simulation of measurements with Hamamatsu R2809U (intrinsic FWHM resolution 20ps) and CFT circuit with pulse selecting threshold level set at low level to accept almost all pulses and different values of delay T_D and attenuation k.

k	0.2	0.2	1
T_D	1.1ns	0.6ns	0.6ns
FWHM	50ps	29ps	24ps

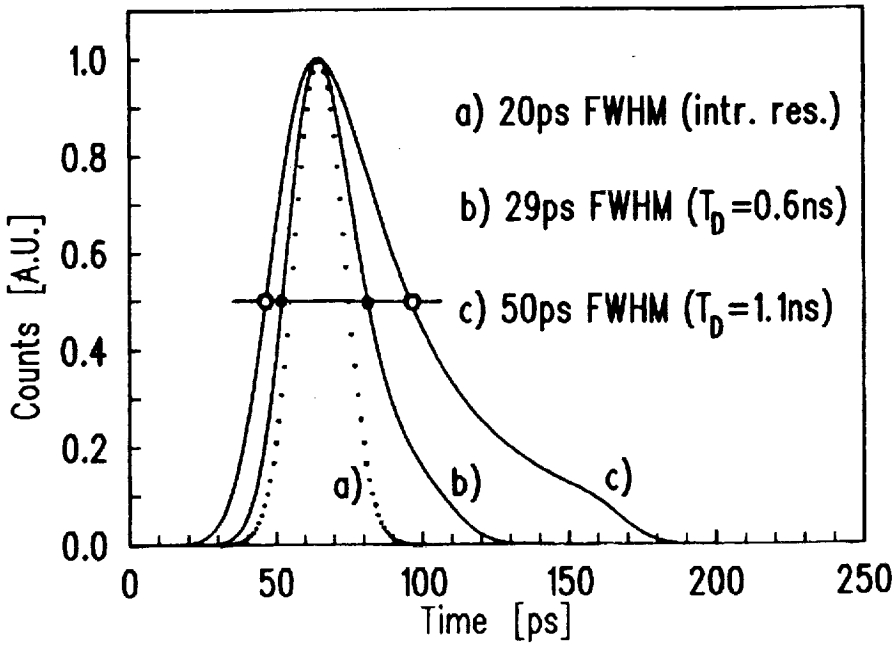

Fig.9 Computed time resolution curves, including effects of amplitude dependent time walk in CFT, for measurement set-ups with MCP model R2809U, fast amplifier HP8447F, CFT pulse selecting threshold set at low level to accept almost all pulses. The dotted curve is the assumed intrinsic resolution curve, having 20ps FWHM. Two cases are reported: a) CFT with the usual attenuation k=0.2 and delay corresponding to minimal external cable T_D=1.1ns (wider curve, 50ps FWHM); b) CFT with k=0.2, but with shorter delay T_D=0.6ns (narrower curve, 29ps FWHM).

It is interesting to compare these results to the corresponding ones for the case of the 12micron channel model R1564U. Essentially, the time walk in the corresponding situations is quite similar in the two cases, but its relative effect on the obtainable performance becomes greater as the intrinsic detector is improved. For the R2809U the actual resolution with the short delay of 0.6ns is remarkable, but 45% wider than the intrinsic one. With delay 0.6ns and no attenuation (k=1) the obtainable resolution is still 20% wider than the intrinsic. Some experimental results point out that the detector intrinsic resolution may be even better than 20ps, possibly 10ps or better.The time walk effect should therefore be considered a major limitation to the ultimate obtainable resolution.

We conclude that there is considerable margin for further improvement. Greater care should be therefore devoted to the electronic pulse processing, in order to take full advantage of the detector intrinsic resolution. It is advisable to work with CFT having k=1 and even shorter delay, possibly with specially studied filter-amplifier input stage. Designing new CFT circuits specially devised for photon counting may be rewarding.

IV. ACTIVE-QUENCHING CIRCUITS FOR SPADs

In early studies on SPADs, the bias arrangement used for the device operation was the so-called *passive quenching* circuit (PQC) [4,5,11]. This circuit employs a high load resistor (in the 100k range) in order to force the diode voltage V to drop down near to the breakdown voltage V_B after each avalanche triggering. This quenches the avalanche. The diode voltage is then slowly restored to the bias voltage, since the diode capacitance is recharged by the small current flowing through the high value resistor. The voltage recovery takes at least some microseconds. A photon can arrive during the recovery from a previous avalanche pulse and trigger the avalanche when the SPAD voltage is at some intermediate level, randomly placed between the breakdown voltage and the correct bias voltage. This has a twofold detrimental effect on the timing performance. First, the intrinsic time resolution of SPADs is strongly reduced as the excess bias voltage $V-V_B$ actually applied to the diode is reduced. Second, at lower excess bias voltage the avalanche current pulse not only has smaller amplitude, but also slower risetime. This causes a walk of the triggering time of the following timing circuit, which is not properly corrected even employing a CFT (CFTs require constant pulse shape to work properly). The intrinsic performance of SPADs can be exploited working with PQCs only in cases where the probability of such events (photons arriving during a recovery from a previous avalanche pulse) is very low, that is, where the rate of repetition of pulses is very low, at best a few kHz. Note that this limitation applies to the total rate of pulses, that is, to the sum of the dark count rate of the SPAD plus the detected photon rate, including unwanted background light.

A partial remedy to such limitations is to apply a pulsed bias voltage to simple passive circuit arrangements, for obtaining a gated operation of SPADs. The additional voltage pulse can be superimposed to the dc bias either by ac coupling or by dc coupling. The ac coupling is very simply implemented, by connecting the junction between SPAD and load resistor to a low-impedance fast pulse generator through a suitable capacitor. The dc coupled gate is obtained by employing a smaller load resistor, typically 1kOhm or less, and applying directly to it the sum of the dc bias voltage plus a pulsed additional voltage. The detailed analysis [18] of such circuit arrangements, however, points out that i) it is possible to detect not more than one photon in the gate interval ii) other specific limitations are associated to each circuit arrangement. With ac coupling, the repetition rate of gating pulses must be low. With dc coupling, the power dissipated in the SPAD may become excessive. In summary, it can be concluded that pulsed-bias passive circuits are of practical interest only for working with gate intervals having short duration, typically below 100ns, and low or moderate repetition rate.

In the early stage of development of SPADs, it became therefore clear that more sophisticated circuits were necessary, in order to fully exploit the device performance. Active quenching circuits (AQCs) were thus conceived and developed in our laboratory [6,12,13]. Essentially, the AQC performs the following operations.

i) It senses the onset of the avalanche current.

ii) It generates an output pulse, synchronous with the avalanche, with the least possible jitter.

iii) It forces the bias voltage of the diode to drop as swiftly as possible below the breakdown voltage. It must therefore have a low-impedance output driver, capable of driving efficiently the capacitance associated to SPADs and connections.

iv) Finally, it restores the initial bias after a well-controlled hold-off time, so that the diode is again ready to detect a subsequent photon. Also the reset transition must be as fast as possible, in order to reduce as far as possible the probability that a photon may arrive during the recovery of the diode voltage, with an associated degradation of the time resolution (see above).

Various problems are met for obtaining a correct AQC operation and severe requirements have to be fulfilled in order to fully exploit the available SPAD performance. A peculiar problem is caused by the large amplitude difference between the avalanche pulse generated by the SPAD and the much larger quenching pulses applied to it and reflected back at the circuit input. The AQC should be sensitive to avalanche pulses of less than 1mA (<50mV over 50), while quenching pulses have amplitude of several Volts (up to 50V in our circuits i.e. 1000 times higher than the avalanche pulse). Unless special precautions are taken in the circuit design, the AQC can be retriggered by the quenching pulse and either be latched in the triggered state or break into a self-sustaining oscillation. Another important requirement is to keep as short as possible the time from the avalanche onset to its quenching. The reason for this is twofold. First, it minimizes afterpulsing effects due to deep levels in the diode junction, acting as charge carrier traps. The charge trapped in deep levels is indeed proportional to the total avalanche charge flowing through the junction. Second, with high voltage devices, the power dissipated in the avalanching state can be fairly high and produce a remarkable variation of the device temperature, with associated variation of the breakdown voltage and of other device parameters. On the other hand, minimizing the delay between the onset of the avalanche and the arrival of the quenching voltage pulse to the SPAD is sometimes conflicting with the requirement of operating the SPAD remote from the AQC, as necessary, for instance, in order to operate a SPAD with cryogenic cooling. For a remotely operated diode, the duration of the avalanche current is inherently increased by twice the transit time in the connecting cables; that is, by 10ns per meter of connecting coaxial cable. The hold-off time must have accurately controlled duration, in order to have a well defined and controlled deadtime (avalanche time plus hold-off time). The actual value of the hold-off time can be very short, a few tens of nanoseconds or less, in cases where photons have to be counted at high rate, up to 10 MHz or more. However, in cases where photon arrival times must be accurately measured and the dark count rate must be minimized, somewhat longer hold-off time have to be employed, typically a few hundred nanoseconds, in order to avoid SPAD retriggering due to the delayed release of trapped carriers. For SPAD devices working with high excess bias voltage $V-V_B$, the quenching pulse must have large amplitude, up to 50V and possibly more, so that the transition times are limited by the slew rate attained by the quenching driver. Finally, in order to exploit the time resolution of the fastest SPADs, the jitter between the avalanche onset and the output pulse should be much lower than the intrinsic resolution of the detector, that is, it must be limited to a few picoseconds. This means that the input stage of the circuit should be designed for minimum noise. Obviously, it is not possible to fulfil at best all these requirements at the same time. Different AQC parameter setting, or even different AQC models are therefore employed for optimizing the most important performance in different cases. In our laboratory, various generations of AQC have been developed, starting in 1975 from the earliest simple model [17] and progressively increasing the performance by a steady evolution of the design [4,6,12,13,18]. Such evolution has produced an AQC design that provides remarkable flexibility for different applications and can work a SPAD in remote position (connected

by a coaxial cable) exploiting at best the device performance. With a minimum-noise input stage, the circuit has intrinsic timing jitter well below 10ps FWHM; it can work at more than 10MHz repetition rate and, with suitably designed output quenching driver, it can provide a 50V quenching pulse.

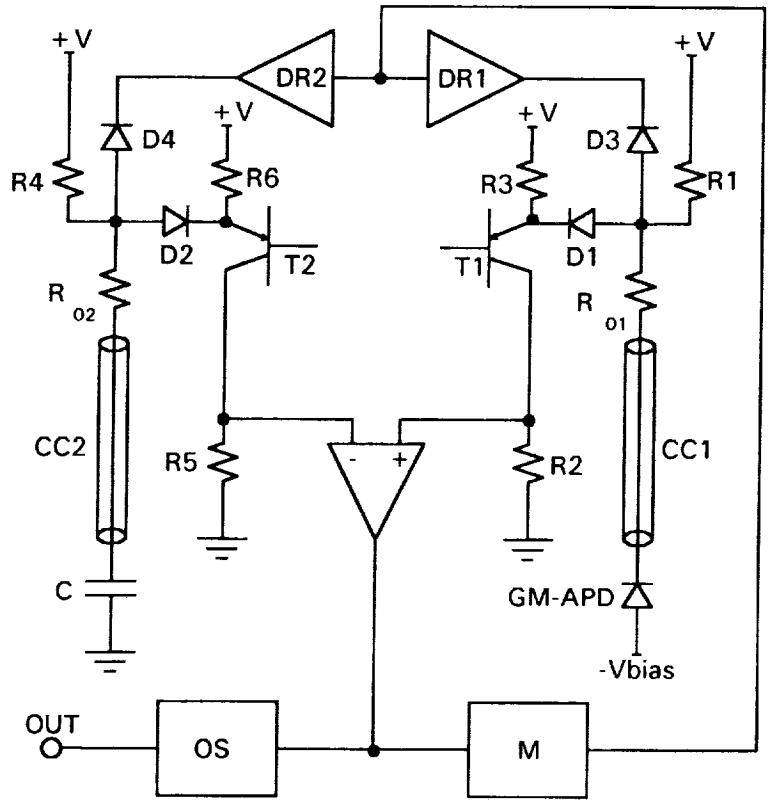

Fig.10 Simplified block diagram of the most recent AQC generation, described in the text. Note the symmetry of the circuit.

The circuit design approach, covered by international patents [13], will be here illustrated making reference to the simplified circuit diagram in Fig.7. The figure illustrates the high degree of symmetry of the circuit. Thanks to this symmetry, the fast comparator can have a low threshold level, as required to sense the avalanche pulse, and at the same time be insensitive to the quenching pulse, avoiding spurious retriggering (see above). The avalanche pulse is applied asymmetrically (only to the non-inverting terminal of the comparator): it is therefore a differential-mode signal that triggers the comparator. The quenching voltage pulse is symmetrically applied (to both input terminals of the comparator), so that it is a common mode signal and does not trigger the comparator. In order to obtain this, the symmetry should be maintained also in the voltage transients, including reflections, overshoots or ringing caused by the load connected to each circuit input. It is therefore necessary to connect to the non-inverting input of the AQC a passive load, in order to match that given by the SPAD, connected to the other input. This normally means a simple capacitor C, with capacitance value equal to that of the SPAD (a trimmer capacitor of a few pF is normally employed). In case of remote SPAD operation, care must be taken of maintaining the

symmetry by connecting with coaxial cables (CC1 and CC2) having equal length the SPAD and the matched load. The cables are terminated at the circuit inputs through the resistors R01 and R02. The avalanche signal flows through the diode D1 and the common-base transistor T1, reaching the comparator input. The role of T1 is twofold. First it establishes a low impedance input, to which the cable-matching resistor R01 is connected. Second, it provides amplification of the voltage signal fed to the comparator, which is important for minimizing the intrinsic time-jitter of the circuit, due to its electronic noise. In stationary conditions, diodes D1 (and D2) is conducting, diodes D3 (and D4) is instead not conducting. When an avalanche is triggered, the current is sensed by the comparator, whose output changes state. This signal triggers a monostable multivibrator, M, that sets the duration of the quenching pulse. In order to obtain the desired amplitude of the quenching signal, a suitable voltage driver stage, DR, is employed. When the quenching voltage pulse is applied by DR, the voltage through D3 becomes direct, the diode conducts and the quenching pulse reaches the photodiode. At the same time the diode D1 is driven to reverse bias condition and prevents the high voltage pulse (up to 50V) from reaching T1. If diode D1 were not present, the entire quenching pulse would be applied to the base-emitter junction of T1, causing it to break down. Another important effect of the two diodes is to break the positive feedback loop of the circuit, thus reducing the risk of oscillations. The AQC output pulse is derived from the comparator, through an output stage OS.

By employing this kind of AQCs, the timing performance of SPADs has been verified down to 20ps FWHM [5]. The flexibility and performance of the circuit have been extensively tested in many different experiments, carried out in a wide variety of conditions, with SPAD working either embedded in the circuit or remote from it and operating in different ambients over a wide range of device temperatures, including cryogenic cooling.

ACKNOWLEDGMENTS

Work supported in part by ASI (Italian Space Agency), CNR (Italian National Research Council) and MURST (Italian Ministry of University and Research). The author wish to thank N.Carbone and S.Masci for their technical support in the development of Active Quenching Circuits.

REFERENCES

1 I.Yamazaki, M.Tamai, H.Kume, H.Tsuchiya and K.Oba, Rev. Sci. Instrum **56**, 1187 (1985).

2 D.Bebelaar, Rev.Sci.Instrum. **57**, 1116 (1986).

3 H.Kume, K.Koyama, K.Nakatsugawa, S.Suzuki, and D.Fatlowitz, Appl. Opt. **27**, 1170 (1988).

4 S.Cova, G.Ripamonti and A.Lacaita, Nucl. Instrum. Methods **A253**, 482 (1987).

5 S.Cova, A.Lacaita, M.Ghioni, G.Ripamonti, T.A.Louis, Rev. Sci. Instrum. **60**, 1104 (1989).

6 S.Cova, A.Longoni, and A.Andreoni, Rev. Sci. Instrum. **52**, 408 (1981).

7 A.Lacaita, S.Cova, M.Ghioni, Rev. Sci. Instrum. **59**, 1115 (1988).

8 S.Cova, M.Ghioni, and F.Zappa, Rev. Sci. Instrum. **62**, 2596 (1991).

9 S.Cova and G.Ripamonti, Rev. Sci. Instrum. **61**, 1072 (1990).

10 S.Cova, M.Ghioni, F.Zappa, and A.Lacaita, Rev. Sci. Instrum. to be published.

11 R.H.Haitz, J. Appl. Phys. **35**, 1370 (1964); **36**, 3123 (1965).

12 S.Cova, A.Longoni, and G.Ripamonti, IEEE Trans. Nucl. Sci. **NS-29**, 599 (1982).

13 S.Cova, US.Patent #4,963,727, Italian patent 22367A/88. Industrial licence to SILENA S.p.A. Milano (Italy).

14 P.W.Nicholson, *Nuclear Electronics* p. 259 (J.Wiley, NewYork, 1974).

15 See e.g. Instruction Manual of ORTEC Mod.583 or Tennelec Mod.454.

16 Hamamatsu Photonics K.K., Hamamatsu City, Japan, and Hamamatsu Corp., Bridgewater, N.J. U.S.A., MCP-PMTs data sheets and Technical Informations No. ET-03/Oct.1987.

17 P.Antognetti, S.Cova and A.Longoni, Proc. Ispra Nucl. Electron. Symp. 1975, Euratom Publication **EUR 537e**, 453 (1975)

18 S.Cova et al. Rev. Sci. Instrum. to be published